

A Set of 200 Musical Stimuli Varying in Balance, Contour, Symmetry, and Complexity: Behavioral and Computational Assessments

Ana Clemente^{1*}, Manel Vila-Vidal², Marcus T. Pearce^{3,4}, Germán Aguiló⁵, Guido Corradi^{1,6}, and Marcos Nadal¹

¹ *Human Evolution and Cognition Research Group, Institute for Cross-Disciplinary Physics and Complex Systems, Associated Unit to CSIC, University of the Balearic Islands, Palma, Spain*

² *Center for Brain and Cognition, Universitat Pompeu Fabra, Barcelona, Spain*

³ *School of Electronic Engineering & Computer Science, Queen Mary University of London, UK*

⁴ *Centre for Music in the Brain, Department of Clinical Medicine, Aarhus University, Denmark*

⁵ *IES Felanitx, Spain.*

⁶ *Department of Psychology, Faculty of Education and Health, University Camilo José Cela, Madrid, Spain*

* Corresponding author: Ana Clemente, Department of Psychology, University of the Balearic Islands, Crta Valldemossa km 7.5 Palma de Mallorca 07122 Spain. Tel. +34645091773. E-mail: ana.c.magan@gmail.com

The project leading to these results has received funding from “La Caixa” Foundation (ID 100010434) under agreements LCF/BQ/ES17/11600021 and LCF/BQ/DE17/11600022, and from the Spanish *Ministerio de Economía, Industria y Competitividad* with grant PSI2016- 77327-P Authors’ contributions: AC created the stimuli and wrote the manuscript; AC and MV designed the computational measures; MV formalized, implemented, and wrote the measures; AC and MN designed the research, discussed the stimuli, and analyzed the data; AC, GC, GA, and MN contributed to the behavioral assessment; AC, MV, MP, and MN compared and discussed the measures, and revised the manuscript. All authors reported no conflicts of interest and approved the manuscript.

Abstract

We present a novel set of 200 Western tonal musical stimuli (MUST) to be used in research on perception and appreciation of music. It consists of four subsets of 50 stimuli varying in balance, contour, symmetry, or complexity. All are 4 s long and designed to be musically appealing and experimentally controlled. We assessed them behaviorally and computationally. The behavioral assessment (Study 1) aimed to determine whether musically untrained participants could identify variations in each attribute. Forty-three participants rated the stimuli in each subset on the corresponding attribute. We found that inter-rater reliability was high and that the ratings mirrored the design features well. Participants' ratings also served to create an abridged set of 24 stimuli per subset. The computational assessment (Study 2) required the development of a specific battery of computational measures describing the structural properties of each stimulus. We distilled nonredundant composite measures for each attribute and examined whether they predicted participants' ratings. Our results show that the composite measures indeed predicted participants' ratings. Moreover, the composite complexity measure predicted complexity ratings at least as well as existing models of musical complexity. We conclude that the four subsets are suitable for use in studies that require presenting participants with short musical motifs varying in balance, contour, symmetry, or complexity, and that the stimuli and the computational measures are valuable resources for research in music psychology, empirical aesthetics, music information retrieval, and musicology. The MUST set and MATLAB toolbox codifying the computational measures are freely available at osf.io/bfxz7.

Keywords: music, aesthetics, MIR, balance, contour, symmetry, complexity

Introduction

Valuing objects is crucial for making decisions, comparing and choosing among alternatives, and prioritizing actions (Berridge & Kringelbach, 2013; Kringelbach, & Berridge, 2009; Levy & Glimcher, 2012). Music is ideally suited for studying evaluative judgments, for three reasons: First, it is a good example of a cultural product whose appreciation relies on basic and general valuation systems (Mallik, Chandra, & Levitin, 2017; Salimpoor & Zatorre, 2013; Shepard, 1982; Trehub & Hannon, 2006). Second, music combines many features of sound to produce virtually unlimited works that vary across composers, styles, times, and cultures (Cross, 2006; Rohrmeier, Zuidema, Wiggins, & Scharff, 2015; Trainor & Unrau, 2011). Finally, people place a high personal value on music (Nieminen, Istók, Brattico, Tervaniemi, & Huotilainen, 2011): they use it to regulate their emotions (Thoma, Ryf, Mohiyeddini, Ehlert, & Nater, 2012) and to enhance the cohesion and coordination in groups (Dissanayake, 2008; Savage, Brown, Sakai, & Currie, 2015), and they are willing to invest time, effort, and money in recorded and live performances (Huron, 2003; Müllensiefen, Gingras, Musil, & Stewart, 2014).

The valuation of music involves the interaction of modality-specific and modality-general attributes (Marin, Lampatz, Wandl, & Leder, 2016; Marin & Leder, 2013; Purwins et al., 2008). Its aesthetic appreciation depends on many factors, including familiarity, perceived complexity, and predictability (Brattico & Pearce, 2013; Edmonston, 1969; Heyduk, 1975; Koelsch, Vuust, & Friston, 2018; Payne, 1980; Pereira et al., 2011; Van den Bosch, Salimpoor, & Zatorre, 2013), which also mediate the valuation of visual stimuli, from architecture to design and

art, (De Lange, Heilbron, & Kok, 2018; Forsythe, Mulhern, & Sawey, 2008; Forsythe, Nadal, Sheehy, Cela-Conde, & Sawey, 2011; Madison & Schiölde, 2017; Tinio & Leder, 2009). Aside from the roles of these factors, however, little is known about the extent to which the valuation of musical and visual objects relies on common attributes. With few exceptions (e.g., complexity in Marin & Leder, 2013), a direct examination of their influence on the valuation of music and visual stimuli has been prevented by the absence of materials comparable across modalities.

In this paper, our goal was to facilitate research on modality-general attributes and domain-general processes in the valuation of music by (1) creating a set of musical stimuli (MUST) suitable for studying modality-general attributes in the valuation of music; (2) assessing the stimulus set behaviorally and computationally; (3) analyzing how both kinds of assessments relate to each other, to stimulus design features, and to existing measures of complexity; and (4) making the MUST set and computational measures available to other researchers through the Open Science Framework (OSF) at osf.io/bfxz7. We designed the set and computational measures to be useful in many fields, including empirical aesthetics, musicology, music psychology, and music information retrieval.

We focused on four attributes: balance, contour, symmetry, and complexity. Their influence on the valuation of visual stimuli is well tested (Gartus & Leder, 2017; Gómez-Puerto, Munar, & Nadal, 2015; Jakesch & Leder, 2015; Locher, Gray, & Nodine, 1996; Palumbo & Bertamini, 2016; Tinio & Leder, 2009; Van Geert & Wagemans, 2019; Vartanian et al., 2018; Wilson & Chatterjee, 2005). For instance, research in empirical aesthetics indicates that

people generally prefer objects and designs that are symmetric (Jacobsen & Höfel, 2002; Gattus & Leder, 2013), complex (Nadal, Munar, Marty, & Cela-Conde, 2010; Machado et al., 2015), balanced (Wilson & Chatterjee, 2005), and curved (Bertamini, Palumbo, Gheorghes, & Galatsidas, 2016; Corradi, Chuquichambi, Barrada, Clemente, & Nadal, 2020). Most of these preferences seem to transcend boundaries of culture (Che, Sun, Gallardo, & Nadal, 2018) and even species (Munar, Gómez-Puerto, Call, & Nadal, 2015).

The effects of these attributes on evaluative judgments are not confined to the visual domain. Evaluative judgments of music are also influenced by contour (e.g., Gerardi & Gerken, 1995; Schmuckler, 2015; Thorpe, 1986; Trehub, Bull, & Thorpe, 1984), symmetry (e.g., Balch, 1981; Bianchi, Burro, Pezzola, & Savardi, 2017; Krumhansl, Sandell, & Sergeant, 1987; Mongoven & Carbon, 2017), complexity (e.g., Marin & Leder, 2013; Pressing, 1999; Steck & Machotka, 1975; Streich, 2007), balance and proportion (Juslin, 2013; Winner, Rosenblatt, Windmueller, Davidson, & Gardner, 1986), as accounted for by a large number of musicological and music-theoretical studies (e.g., Cook, 1987; Grey, 1988) and treatises on form (e.g., Caplin, Hepokoski, & Webster, 2010; Leichtentritt, 1911) and composition (e.g., Schoenberg, A., 1967). Could the fact that balance, contour, symmetry, and complexity influence evaluative judgments in the visual and musical domains owe to cross-modal processes? Testing this intriguing possibility requires, however, materials that are directly comparable, analogous in specific dimensions in the auditory and visual modalities.

We intended our stimuli to be both musically appealing and experimentally controlled. Excerpts from the existing repertoire (e.g., Marin & Leder, 2013;

Egermann, Pearce, Wiggins, & McAdams; 2013; Gingras et al., 2016), have the advantage of being naturalistic, but also the drawback that some might be more familiar than others, have different duration, and include other sources of uncontrolled variability. Conversely, controlled sequences of synthesized sounds can minimize extraneous variables (e.g., Shmulevich & Povel, 2000; Steck & Machotka, 1975), but they also reduce musical appeal and ecological validity. We therefore chose to compose motifs that combine the musical appeal of genuine musical excerpts with the experimental control of synthesized sequences.

Once the stimuli were composed, we subjected them to two assessments. First, we conducted a behavioral experiment (Study 1) to determine whether the design parameters we manipulated to produce variations in balance, contour, symmetry, and complexity translated into perceived variations in each of these attributes by musically untrained participants. Based on the results of this experiment, we created an abridged set of stimuli to be used more efficiently in experimental settings. Second, we developed several computational measures for each parameter manipulated to compose the stimuli (Study 2). These computational measures served (i) to describe each motif in terms of structural properties, (ii) to derive nonredundant composite measures for each attribute (balance, contour, symmetry, and complexity), (iii) to ascertain which of the composite measures, or combination thereof, explain participants' assessments of the stimuli attributes in the behavioral experiment, and (iv) to compare the explanatory adequacy of our composite measures of complexity with other objective methods for computing musical complexity.

Design of the Musical Stimuli

The MUST set consists of 200 original musical motifs composed by the first author—an accomplished professional musician with broad compositional and performing experience—using Finale 2012 (MakeMusic Coda Music Technologies), and comprising four subsets of 50 stimuli that vary in terms of a specific attribute: Balance, Contour, Symmetry, and Complexity. Four additional motifs were composed for each subset to be used as examples while giving experimental instructions.

The motifs in the MUST Balance subset capture and translate into music the variation in balance among the visual stimuli in Wilson and Chatterjee's (2005) set. This set consists of diverse arrangements of seven hexagons or circles of distinct sizes. These stimuli were created to vary in balance, measured as the average of eight symmetry components over the axes of the stimuli (Figure 1, first column). The motifs in the Contour subset reflect the kind of variation between the curved and sharp contours of Bertamini et al.'s (2016) visual stimuli. These stimuli were designed as closed black figures based on circles, ovals, or lobed ovals, and matched in the number of vertices. Half of them had curved contours, and the other half had sharp-angled contours (Figure 1, second column). The musical motifs in the Symmetry and Complexity MUST subsets were composed to capture the variation in symmetry and complexity in Jacobsen and Höfel's (2002) set of visual designs. This set consists of a series of images of solid black circles with a centered white square containing triangles that are combined to form designs of varying complexity and symmetry. Half of the configurations are symmetric, and the other half, asymmetric, and the stimuli in both halves match

for different degrees of complexity, corresponding to the number of constituent elements (Figure 1, third and fourth columns). Unlike Jacobsen and Höfel (2002), who developed visual designs varying in both symmetry and complexity, we present a subset varying in complexity and a separate one varying in symmetry.

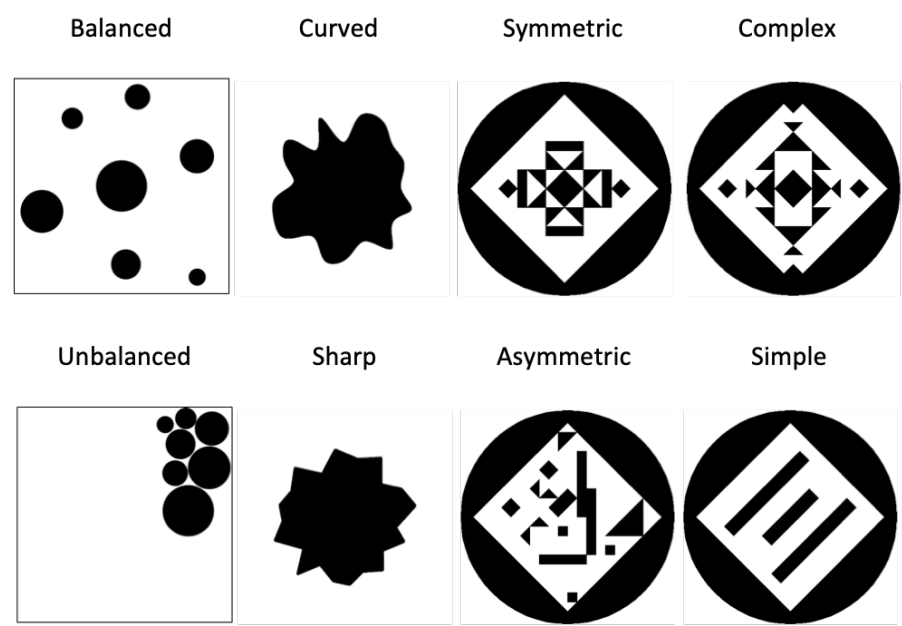


Figure 1. Examples of visual stimuli designed by Wilson & Chatterjee (2005) for balance; Bertamini, Palumbo, Gheorghes, & Galatsidas (2016) for contour; and Jacobsen & Höfel (2002) for symmetry and complexity.

The composer used her musical and artistic expertise to manipulate specific musical parameters to generate variation within each target attribute: balance, contour, symmetry, and complexity (Table 1). The compositional process also aimed to make the set coherent, and the stimuli comparable across sensory modalities and equivalent in musical attributes.

Table 1. *Summary of Parameters used to Design the Musical Stimuli in each Subset*

Attribute	Parameter	Feature	
		<i>Balanced</i>	<i>Unbalanced</i>
Balance	Distribution of elements/ events	Regular	Irregular
	Climax position	Centered	Skewed
	Tension	Progressive	Unprepared
Contour		<i>Smooth</i>	<i>Jagged</i>
	Intervals	Only small (\leq fourths)	Large ($>$ fourths) & small
	Durations	Progressive, small changes	Sudden, large changes
Symmetry		Symmetric	Asymmetric
	Vertical mirror structure	Yes	No
Complexity		<i>Simpler</i>	<i>More complex</i>
	Number of elements/events	Few	Many
	Variety of elements/events	Low	High
	Predictability	High	Low

Mirroring the sets of visual images described above, the motifs in the Complexity subset vary along a continuum (from *simpler* to *more complex*). In contrast, those in the other three sets belong to one of two poles: *balanced* vs. *unbalanced* (Balance), *smooth* vs. *jagged* (Contour), and *symmetric* vs. *asymmetric* (Symmetry) (see Figure 2 for examples of the scores, and Table 1 for the parameters used to design the stimuli). For the Balance, Contour, and Symmetry subsets, the stimuli were designed to achieve high between-pole and low within-pole variation in the target parameters, while minimizing variation in the other parameters. Because timbre and intensity are constant across all stimuli, variations in the four attributes were created using pitch, rhythm, and harmonic implication.

Balance subset. Stimuli vary in their equilibrium, as applied both to the distribution of notes throughout the motif and to the distance of the tensional peak from the central time point. Melodic and harmonic tension contribute to the climax and consequently to balance, but for such brief and constrained stimuli, it stands to reason that they play a weaker role than the distribution of notes in time. A motif is *balanced* if its notes are uniformly distributed with relation to a central climax (or center of mass, in analogy to physical gravity). A motif is *unbalanced* if most notes accumulate at either the beginning or ending.

Contour subset. Stimuli differ in terms of interval size and rhythmic change, leading to differences in the profile of their melodic line. Although contour may also refer to the direction of melodic movement (i.e., rising, falling, or constant pitch intervals regardless of their size), we define it as melodic shape or configuration, thus determined by interval size and duration (or onset) ratios. Therefore, for the *smooth* motifs, we used only small intervals (\leq fourths, predominantly seconds) and rhythms in which successive note durations change very little, while *jagged* motifs included large intervals (> 4 ths) and sudden rhythmic shifts.

Symmetry subset. Stimuli differ to the extent they are symmetrical around a central vertical axis. In *symmetric* motifs, the second half is a literal retrograde repetition of the first half. They thus have a mirror reflection structure—e.g., A(B)A, ABC(C)BA. The only exception to strict symmetry is that the duration of the first and last notes may not be equal because of notational constraints. In *asymmetric* motifs, there is no such retrograde repetition.

Complexity subset. Stimuli vary in the number, variety, and predictability of their elements or events. *More complex* motifs have many notes varying widely in duration, pitch interval size, and register. Conversely, *simpler* motifs are characterized by a small number of highly predictable notes with repeated uncomplicated patterns.



Figure 2. Musical stimuli sample scores in each subset, all to be played in $\text{♩} = 120$ (i.e., quarter note at 120 bpm).

We strove to minimize variation in all attributes other than the intended one, even though we expected some inter-correlations between the parameters defining different attributes. For instance, all other parameters being equal, *symmetric* patterns will be judged as simpler than *asymmetric* designs, both in the visual and the auditory modalities, as they imply redundancy by definition. This is why all stimuli in the Complexity subset are symmetric, all included in Contour

and Balance are asymmetric, and all stimuli in the Symmetry subset have medium to low complexity (as complexity hampers the perception of symmetry; Mongoven & Carbon, 2017). We obtained estimates of the file sizes of the musical motifs using lossless compression format FLAC (Free Lossless Audio Codec) to uncompressed WAV (Waveform Audio File Format) files, for it appears to be a good approximation of complexity ratings of musical stimuli (Marin & Leder, 2013). This enabled us to ensure that the *asymmetric* and *symmetric* poles of the Symmetry subset did not differ significantly in terms of complexity ($t_{(48)} = 1.595, p = .117$) as assessed by FLAC compression. Just like visual curves imply more information than polygons, the pitch entropy is higher by definition in the *jagged* than in the *smooth* stimuli. However, the *t*-tests revealed no significant differences between the poles of the Contour subset ($t_{(48)} = 2.007, p = .050$). In contrast, the FLAC compression sizes of the *unbalanced* motifs were, overall, significantly larger than those of the *balanced* ones ($t_{(48)} = 6.555, p < .001$), probably because self-similarity may be higher in balanced designs. Furthermore, symmetry in the visual and music domains can be regarded as an extreme form of balance. Therefore, all motifs except the *unbalanced* were composed with a high degree of balance. Finally, all except those in the Contour subset possessed medium contours (not too jagged, not too smooth).

Short monophonic melodies are the musical analogues to the abstract visual patterns in the visual sets. Although musical pieces are often polyphonic, we retained the underlying harmony in our motifs, together with the factors related to the stimulus that may define the attributes in both short monophonic and long polyphonic music. To avoid harmony being unduly affected by the

manipulations, we carefully used simple harmonic sequences and rhythmic figures, thereby maintaining the musical structure and style similar for both poles in the Balance, Contour, and Symmetry subsets. Finally, tessiture and tempi were compensated within subsets and never extreme. The fastest tempo is 180 bpm, and the pitch range spans from C₂ to C₆ (provided A₄ = 440 Hz), approximately the human vocal range.

All stimuli were composed using the same musical idiom, including language and style (Western tonal-functional), key (C-Major), texture (monophonic), timbre (piano-like; Garritan Sound Library for Finale, MakeMusic), duration (4 s), overall and instantaneous loudness (no changes in musical dynamics or spatial cues), and other acoustical properties (i.e., expressive performance and recording inconsistencies and variability are nonexistent). A length of 4 s seems optimal for experimental settings where visual correspondence is of relevance because it does not imply an excessive working memory load and approximates presentation times of images in studies of visual aesthetics, allowing comparisons between auditory and visual research findings. Moreover, nonmusicians' short-term memory for music is thought to span about 3-5 s (Schaal, Banissy, & Lange, 2015; Snyder & Snyder, 2000), and the perception of musical symmetry is optimal within this duration (Mongoven & Carbon, 2017; Petrović, Ačić, & Milanković, 2017).

Study 1: Behavioral Assessment of Musical Stimuli

Method

Participants.

Forty-three self-reported nonmusicians (none of whom had ever received higher education in music or was a professional musician; see full questionnaire in Appendix A, Supplementary Materials) aged 18–55 years ($M = 29.31$, $SD = 10.56$, 24 female, 18 male, one not reported) took part in the study. All gave informed consent before participating and reported normal or corrected-to-normal vision and hearing, and no cognitive impairments. Participants were unaware of the purpose of the study, and all procedures followed local ethical guidelines and the Declaration of Helsinki.

Materials.

The stimuli were the 200 motifs described above, and the four example stimuli for each subset, presented in WAV format using Open Sesame (Mathôt, Schreij, & Theeuwes, 2012).

Procedure.

The study was conducted at the Laboratory of Psychology of the University of the Balearic Islands. Each of the 43 participants rated each of the 50 musical motifs in each subset presented as a different experimental block consisting of instructions (available in Appendix A), four examples (two for each pole) to illustrate the instructions, five practice trials, and the experimental task itself. The five stimuli for the practice trials were selected from the 50 in each subset, counterbalanced across participants. Thus, although participants rated 45

stimuli in each subset, all 50 stimuli received ratings. The order of the blocks was also counterbalanced. The order of the 45 stimuli used in the experimental task was randomized individually. All stimuli were presented in sound-attenuated cabins through headphones.

At the beginning of each block, a text introduced and defined the attribute according to its design parameters, and four illustrative examples were played. During the first examples, the participants adjusted headsets and volume to personal comfort levels, which remained unmodified throughout the experiment. They then rated the five practice stimuli under the experimenter's supervision and assistance. After the experimenter had made sure that participants understood the task and all doubts had been resolved, the participants rated the 45 remaining stimuli alone using Likert scales ranging from 1 to 5 and anchored by *very balanced* (1) and *very unbalanced* (5) for Balance, *very smooth* (1) and *very jagged* (5) for Contour, *very symmetric* (1) and *very asymmetric* (5) for Symmetry, and *very simple* (1) and *very complex* (5) for Complexity. The rating scale appeared after each musical motif had ended, and served as a cue for response. The rating was self-paced, and the participants could play each stimulus as many times as they wished before rating it. The procedure was the same for all blocks. After finishing each block, the participants could rest for a moment before going on to the next. A brief questionnaire followed the last block, recording information on demographics, musical education, and general feedback (included in Appendix A). The whole experimental session lasted about 40 minutes, after which the participants were debriefed and thanked.

Data analysis.

This behavioral assessment had two objectives. The first was to determine whether untrained participants perceived variations in the defining attribute for each subset, that is to say, whether stimuli designed to be more complex, for instance, would indeed be perceived and rated by nonmusicians as more complex. To this end, we first assessed inter-rater reliability for each subset using intraclass correlation coefficients ($ICC_{3,k}$; Shrout & Fleiss, 1979). We then conducted Wilcoxon signed-rank tests (given that the Shapiro–Wilk test of normality revealed that several of the distributions were not normal) to determine whether mean ratings for stimuli in each pole in the dichotomous subsets (Balance, Contour, and Symmetry) differed significantly. For the continuous subset (Complexity), we calculated the Spearman correlation between the FLAC file size of each musical motif and its mean rating.

The second aim was to select part of the musical motifs in each subset to assemble an abridged set that could be applied in future studies in a shorter session. We wished to include motifs that participants agreed belonged to the different poles in each subset. Following Nadal et al.'s (2010) method, we calculated the mean and standard deviation of each stimulus' ratings. For each subset, we selected the 12 stimuli with the highest mean rating and the 12 stimuli with the lowest mean rating (those perceived as most *balanced*, *smooth*, *symmetric*, and *simple*, and those perceived as the most *unbalanced*, *jagged*, *asymmetric*, and *complex*), provided the standard deviation of participants' ratings was below the 75th percentile, and that the mean rating placed the motif in the pole it was designed to be in. We thus assembled four subsets containing 24

stimuli each, 12 in each pole, maximizing the difference between and minimizing the difference within levels. This ensured that stimuli represented extreme poles of each dimension and that participants did not disagree on their allocation. Finally, to verify whether the motifs in each pole of each subset of the abridged set actually corresponded to different levels, we compared their mean ratings using Wilcoxon nonparametric tests.

Results

Inter-rater reliability.

The average fixed raters' ICC was high for all subsets: for Balance, $ICC_{3,k} = .94$, 95% CI [.92, .96]; for Contour, $ICC_{3,k} = .97$ [.96, .98]; for Symmetry, $ICC_{3,k} = .84$ [.77, .90]; for Complexity, $ICC_{3,k} = .99$ [.98, .99]. These values show that participants understood the task and judged the stimuli in very similar ways.

Ratings of attributes.

According to the Shapiro–Wilk tests, the mean ratings of each motif were not normally distributed for the Balance ($W = 0.842$, $p < .001$, skew = -0.092, kurtosis = -1.788), Contour ($W = 0.843$, $p < .001$, skew = 0.052, kurtosis = -1.790), and Complexity ($W = 0.85147$, $p < .001$, skew = -0.713, kurtosis = -1.040) subsets, whereas the distribution of Symmetry ratings did not differ significantly from normality ($W = 0.982$, $p = 0.628$).

Participants' ratings corresponded well to the design features of musical motifs in each subset (Figure 3). Wilcoxon tests showed significant differences between the mean ratings of *balanced* ($M = 2.2$, $SD = 0.3$) and *unbalanced* ($M = 3.82$, $SD = 0.18$) motifs in the Balance subset ($W = 0$, $p < .001$), between *jagged*

($M = 3.99$, $SD = 0.32$) and *smooth* ($M = 2.06$, $SD = 0.25$) motifs in the Contour subset ($W = 625$, $p < .001$), and between *asymmetric* ($M = 3.09$, $SD = 0.54$) and *symmetric* ($M = 2.47$, $SD = 0.41$) motifs in the Symmetry subset ($W = 513$, $p < .001$). Spearman correlation analysis indicated a strong relation between the FLAC file size and mean rating for the motifs in the Complexity subset ($r_s = .78$, $p < .001$). In sum, reflecting the design features of the stimuli, participants gave higher unbalance scores to the *unbalanced* stimuli than to the *balanced* stimuli, higher jaggedness scores to *jagged* than to *smooth* stimuli, higher asymmetry scores to *asymmetric* than to *symmetric* stimuli, and higher complexity scores to *complex* than to *simple* stimuli.

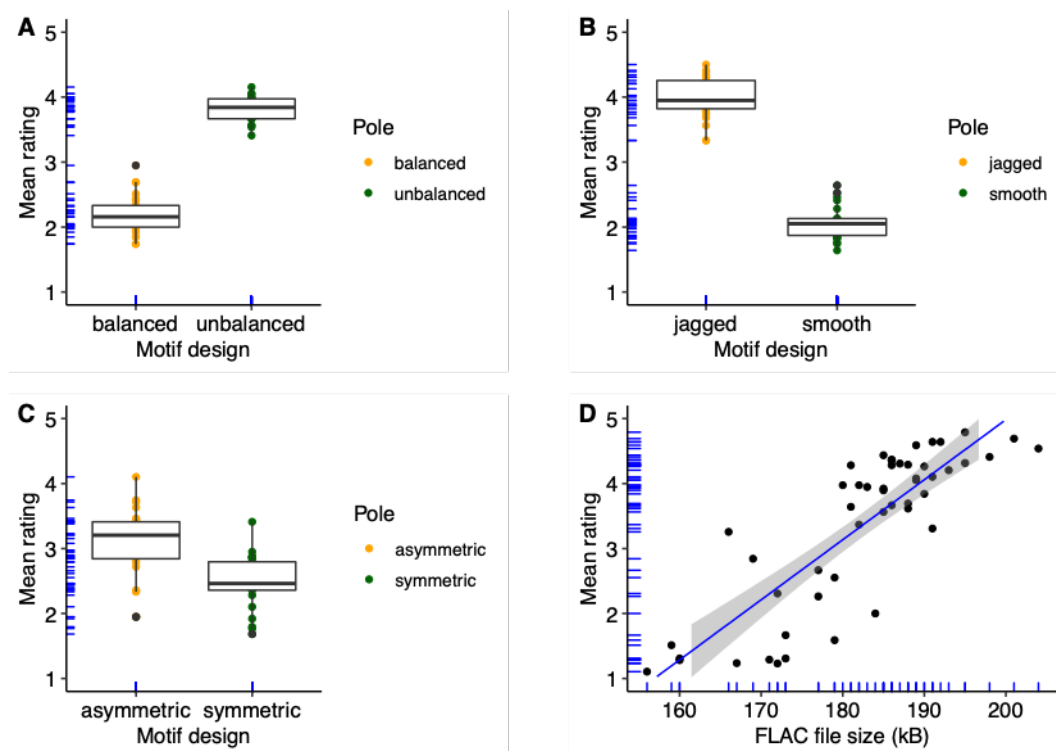


Figure 3. Correspondence between the behavioral assessment and the design of the motifs. Boxplots are used for the discrete subsets of Balance (A), Contour (B), Symmetry (C), and a scatterplot illustrates the continuous subset: Complexity (D). The boxes represent the median, first and third quartiles; whiskers span $Q1 - 1.5 \times IQR$ (interquartile range) to $Q3 + 1.5 \times IQR$. For the Complexity subset (D), the regression line is depicted with its 95% CI (gray ribbon). kB refers to kilobytes. The figure includes rug plots of mean ratings (left), and FLAC file size for the Complexity subset (bottom).

Creation of the abridged set.

Following the procedure described above, we selected the 12 stimuli that received the most extreme ratings of balance and unbalance, smoothness and jaggedness, symmetry and asymmetry, and simplicity and complexity, provided there was no strong disagreement among the raters (Balance $SD_{75th} = 1.40$; Contour $SD_{75th} = 1.26$; Symmetry $SD_{75th} = 1.57$; Complexity $SD_{75th} = 1.01$). We also selected two additional stimuli from each pole of each subset (the next two most extreme items) to be used as practice trials when employing the abridged set. The whole abridged set therefore includes 96 musical motifs representing the extreme poles of balance, contour, symmetry, and complexity, plus 16 practice stimuli. The list is available in Appendix C in the Supplementary Materials.

Figure 4 graphically represents the relation between the mean and the standard deviation of the ratings for each stimulus. The general trend, at least in the Symmetry and Complexity subsets, is for participants to agree more in their ratings of stimuli close to the extreme of the poles, and less in their rating of stimuli far from the poles. Wilcoxon tests indicated that for each of the abridged subsets, the selected stimuli in each pole (filled dots in Figure 4) received significantly different ratings (for each of the four abridged subsets separately, $W = 0, p < .001$). Thus, in the abridged subsets, the rated unbalance for *unbalanced* stimuli ($M = 3.92, SD = 0.1$) was higher than for *balanced* stimuli ($M = 2.01, SD = 0.17$), the rated jaggedness for *jagged* stimuli ($M = 4.16, SD = 0.22$) was higher than for *smooth* stimuli ($M = 1.93, SD = 0.15$), the rated asymmetry for *asymmetric* stimuli ($M = 3.49, SD = 0.27$) was higher than for *symmetric* stimuli

($M = 2.2$, $SD = 0.33$), and the rated complexity was higher for *complex* stimuli ($M = 4.51$, $SD = 0.16$) than for *simple* stimuli ($M = 1.49$, $SD = 0.36$).

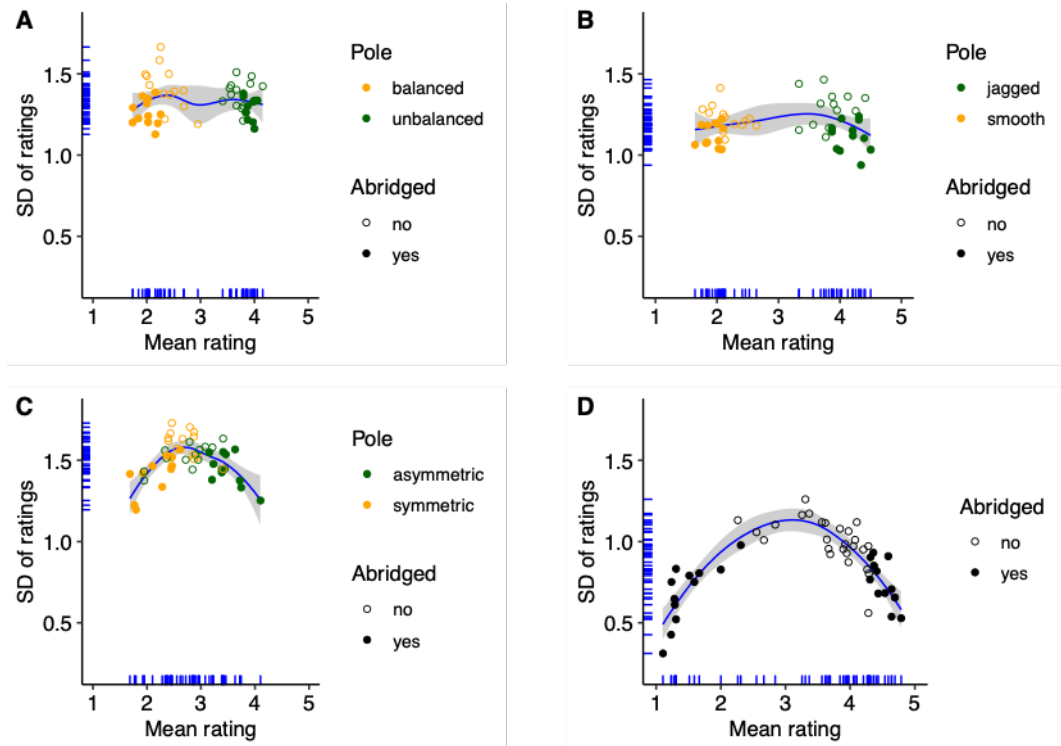


Figure 4. Distribution of means and standard deviations of ratings for each musical motif in each subset: Balance (A), Contour (B), Symmetry (C), and Complexity (D). Filled dots correspond to motifs selected for the abridged set. The figure includes rug plots of the mean (bottom) and the standard deviation (SD) of the ratings (left). Curved lines depict local polynomial regression fitting ($SD \text{ ratings} \sim M \text{ ratings}$), for which the gray ribbon represents the 95% CI.

Discussion

The overarching goal of our research was to facilitate the investigation of modality-general attributes and domain-general processes in the valuation of music (see also Margulis, 2016). To this end, we created four subsets of 50 brief musical motifs varying along a single dimension (balance, contour, symmetry, or complexity) for use in empirical aesthetics, musicology, music psychology, and

other fields. We conducted a behavioral assessment of the stimuli with two aims: First, we wished to determine whether musically untrained participants noticed the variations in each subset, that is whether they could distinguish between the *balanced* and *unbalanced*, *smooth* and *jagged*, *symmetric* and *asymmetric*, and *simpler* and *more complex* motifs. Second, we wished to assemble an abridged version of our four subsets that could be applied in future studies in a shorter session.

The results of the behavioral assessment show that participants were clearly able to distinguish the stimuli with respect to their defining attribute. This means, first, that variations in each of the attributes were readily perceptible to participants, and second, that participants' ratings concurred with the design of the stimuli. The results also revealed a very high inter-rater reliability, suggesting that participants understood the task in a similar way and judged the musical motifs using common criteria. This holds for all subsets, although the differentiation between *symmetric* and *asymmetric* motifs of the Symmetric subset seems to be less apparent than the distinction between the poles of other dichotomous subsets. A plausible explanation is that musical symmetry may require higher memory load and levels of attention than other attributes: one would have to memorize and compare events of the motif several seconds apart and in reversed order with high accuracy to discern whether it is symmetric (Krumhansl et al., 1987; Mongoven & Carbon, 2017). Nevertheless, even though slightly lower, the inter-rater reliability was still high, and while the standard deviations were slightly higher for the Symmetry subset, these values were not excessive, and the mean ratings for each pole were significantly different. Participants' ratings, in sum, reliably mirrored

the design parameters of the musical motifs. We conclude, therefore, that the four subsets are suitable for use in future studies that require presenting participants with short musical motifs varying in balance, contour, symmetry, or complexity.

The presentation of 50 stimuli in each subset might be too long in some studies. We therefore used the ratings provided by our participants to derive an abridged version of each subset, selecting the 24 stimuli that represented the most extreme poles of each attribute, and for which there was no substantial disagreement among raters. As a general trend, the agreement among participants was highest for stimuli close to the extremes. We also selected training stimuli for each attribute. Thus, the complete abridged set contains 96 short musical motifs to be used in future studies, in addition to 16 equivalent training motifs (2 for each pole of each of the 4 attributes): the abridged Balance subset includes 12 clearly *balanced* and 12 clearly *unbalanced* musical motifs, the abridged Contour subset includes 12 clearly *smooth* and 12 clearly *jagged* musical motifs, the abridged Symmetry subset includes 12 clearly *symmetric* and 12 clearly *asymmetric* musical motifs, and the abridged Complexity subset includes 12 clearly *simple* and 12 clearly *complex* musical motifs.

Study 2: Computational Assessment of Musical Stimuli

This study had four main goals: (1) to develop a series of specific computational measures that provide a suitable description of each of the 200 musical motifs in terms of structural properties, (2) to derive nonredundant composite measures for each attribute, (3) to determine which of the composite measures, or combination thereof, explained participants' ratings of each attribute

in Study 1, and (4) to compare the explanatory adequacy of our composite measures of complexity with existing methods. Our aim was, therefore, to find the most complete model integrating the contributions of all parameters manipulated in the design of the stimuli.

Method

Computational measures of musical attributes.

We implemented several basic, conceptually irreducible, compact, and quantitative computational measures of the design parameters of each of the four attributes. Appendix B in the Supplementary Materials describes the measures in detail, and Appendix C presents the values of the computational measures for each stimulus in each corresponding subset.

Higher values correspond to more unbalance, jaggedness, asymmetry, and complexity. The measures were devised to assess each of the attributes in our MUST set, but we expect them to generalize to other stimuli, experimental paradigms, and researchers. A comprehensive description and formulation of the computational measures, together with a rationale for their selection, is presented as Appendix B in the Supplementary Materials. The corresponding functions for MATLAB integrate the MUST toolbox, available at osf.io/bfxz7 and <https://github.com/compaes>.

Balance. As conceived here, balance is related to the distribution of events and the position of the climax in the course of a tensional process. We implemented three measures that capture three different aspects of the global perception of balance based on the distribution of events and the relative positions of each motifs' center of mass and geometric center (Table 2).

Contour. Contour perception is related to the magnitude of changes in pitch and duration. Small changes are perceived as smooth, whereas large changes are perceived as abrupt or jagged. We implemented three measures of intervallic and melodic abruptness, and one measure of rhythmic abruptness (Table 2).

Symmetry. The only form of symmetry considered is vertical mirror reflection: the strict retrogradation of all sounds (pitch and duration) from a central axis. Due to notation restrictions, an adjustment of the last note duration was sometimes needed (to equalize it to the first one). We implemented two measures of this kind of musical symmetry (Table 2).

Complexity. The complexity of the motifs was manipulated by varying the quantity and variety of elements in pitch and duration, resulting in variations in predictability. We implemented one measure of the number of elements, and seven measures that capture different aspects related to the variety of elements and their predictability (Table 2).

Table 2. *Computational Measures of the Parameters Used to Compose Musical Motifs Varying in Balance, Contour, Symmetry, and Complexity*

Attribute	Parameter	Computational measure
Balance	Distribution of elements/events	<i>Bisect unbalance</i> : Equilibrium between the two halves of a stimulus
	Climax position	<i>Center of mass offset</i> : Distance between center of mass and geometric center
	Tension	<i>Event heterogeneity</i> : Heterogeneity in the temporal distribution of events
Contour	Intervals	<i>Average absolute interval</i> : Average absolute pitch interval size
		<i>Melodic abruptness</i> : Average interval size of changes of direction per note
		<i>Durational abruptness</i> : Proportion of the stimulus with changes of direction
	Durations	<i>Rhythmic abruptness</i> : Average ratio of consecutive durations
Symmetry	Vertical mirror structure	<i>Total asymmetry</i> : Direct–retrograde accumulated pitch difference <i>Asymmetry index</i> : Proportion of the stimulus with asymmetries
	Number of elements/events	<i>Event density</i> : Number of note events per time unit
Complexity	Variety of elements/events Predictability	<i>Average local pitch entropy</i> : Average pitch entropy of .25-s sliding windows
		<i>Pitch entropy</i> : Entropy of pitch distribution
		<i>2-tuple pitch entropy</i> : Entropy of 2-tuple pitch distribution
		<i>3-tuple pitch entropy</i> : Entropy of 3-tuple pitch distribution
		<i>2-tuple interval entropy</i> : Entropy of 2-tuple interval distribution
		<i>3-tuple duration entropy</i> : Entropy of 3-tuple duration distribution
		<i>Weighted permutation entropy</i> : Permutation entropy considering the <i>SD</i> of the pitch distribution of each 3-note sequence

Our battery of measures takes advantage of the state of the art in music information research, music cognition, and related fields, while going further in designing new measures. For instance, *event density* and *pitch entropy* are common in existing models of perceived complexity, such as Eerola et al.’s

Expectancy-Violation model (EV; Eerola, 2016). However, Eerola and colleagues based their analysis on pitch classes, whereas we consider absolute pitch, and our measures of entropy of pitches go beyond pitch entropy in considering, for example, the entropy of tuples and intervals (see Appendix B). Some measures include an application of established principles and algorithms correspondingly cited (e.g., Shannon entropy, Parncutt's model), while other measures are entirely original (e.g., Symmetry measures).

To determine whether variation in the parameters pertaining to each attribute actually contribute to variation in that attribute and not—or not significantly—to variation in the other three attributes, we also applied the full battery of measures detailed in Table 2 to the 200 musical motifs. The results indicate that the manipulations of parameters pertaining to any given attribute did not result in notable effects on other attributes. This analysis is reported in Appendix D.

Composite nonredundant measures.

Given that the measures described above capture different aspects (e.g., *melodic abruptness* and *durational abruptness*) of the same attribute (e.g., contour), we expected multiple regression models to contain some redundancy and multicollinearity. Therefore, we conducted four principal components analyses (PCA), one for each attribute, in order to extract nonredundant components for each attribute. We then used these components as predictors of participants' ratings.

Before running the PCA, several tests were conducted to evaluate the adequacy of the data for factor analysis. Bartlett's test of sphericity quantifies the

overall significance of all correlations within the correlation matrix ($p < .050$). The Kaiser-Meyer-Olkin ($KMO > .50$) assesses the sampling adequacy and the strength of the relationships among variables. Values of the determinant of the correlation matrix over 10^{-5} indicate an acceptable amount of multicollinearity in the data set.

Factors were retained following Jolliffe's (eigenvalues > 0.70 ; Jolliffe, 1972) criterion and inspecting the cumulative proportion explained. When extracting more than one factor, *oblimin* rotation was performed, given that factors relating to the same attribute were not entirely orthogonal. We calculated the component scores for each stimulus and treated these as composite computational measures of balance, contour, symmetry, and complexity in the subsequent analyses.

Explaining participants' ratings of musical attributes.

We used linear mixed-effects models (Hox, Moerbeek, & van de Schoot, 2010; Snijders & Bosker, 2012) to analyze the effects of the predictors (the composite computational measures obtained in the PCA) on participants' responses for each subset. They account simultaneously for the between-subject and within-subject effects (Baayen, Davidson, & Bates, 2008), and are thus especially suitable for responses that may vary between individuals and stimuli (Silvia, 2007; Brieber, Nadal, Leder, & Rosenberg, 2014; Cattaneo et al., 2015; Vartanian et al., 2018). We created a model for each subset to assess the predictive power of the components with respect to participants' responses. The structure of all models was the same. We modeled the behavioral ratings of balance, contour, symmetry, and complexity considering the corresponding composite measures,

and their interactions when more than one, as fixed effects. We included random intercepts and slopes for the composite measures, and their interaction when more than one, within participants, following Barr, Levy, Scheepers, and Tily's (2013) recommendation to model the maximal random-effect structure. In addition to avoiding loss of power and reducing type I error, this enhances the possibility of generalizing results to other participants.

Following Aguinis, Gottfredson, and Joo (2013), and considering the nature of our study, we looked for highly influential observations among participants' ratings by inspecting Cook's distance (Cook, 1979). The threshold was set at $4/(N-k-1)$, where N is the number of observations ($N = 43$) and k is the number of explanatory variables.

All analyses were carried out within the R environment for statistical computing, R version 3.5.1. (R Core Team, 2018). We used the *principal()* function in the 'psych' package (Revelle, 2018), the *lmer()* function of the 'lme4' package (Bates, Maechler, Bolker, & Walker, 2015) and the 'lmerTest' package (Kuznetsova, Brockho, & Christensen, 2012) to estimate the p -values for the t -tests based on the Satterthwaite approximation for degrees of freedom, and the 'influence.ME' package (Nieuwenhuis, te Grotenhuis, & Pelzer, 2012). Effect sizes were calculated following Judd, Westfall, and Kenny's (2017) indications.

Comparison with other objective measures of complexity.

We are unaware of other computational measures or models of perceived balance, contour, and symmetry that we could compare with our own. There are, however, several general models of perceived musical complexity, and we compared the performance of these models with the ability of our composite

models to predict participants' complexity ratings. Order is thought to influence the perception of complexity in both domains, as discussed in Nadal et al. (2010) and Van Geert and Wagemans (2019). Besides the number and variety of events, the computational measures within the MUST complexity model (MUST_K) quantify various forms of entropy of sequences of pitches, intervals, and durations, thus accounting for diverse kinds of order and predictability, characteristic of the musical language. These qualities make the MUST_K model suitable for comparison with models such as the expectancy-violation model or the IDyOM. For fair comparison, we only considered complete models developed at the same explanatory level and addressing the same dimension (cf., Marin & Leder, 2013). We selected three models that are suitable for short stimuli, that have been demonstrated to be the best in their respective categories, and that have been validated with Western tonal music:

FLAC compression. Free Lossless Audio Codec (FLAC) is a compression format specific for audio files (Coalson, 2008) that incorporates a linear autoregressive predictor and has been proven a good indicator of perceived musical complexity based on data redundancy (Marin & Leder, 2013). In contrast to generic systems such as ZIP, special attention is placed on the temporal organization of structures (Robinson, 1994). We employed the default settings at an online FLAC converter (<https://audio.online-convert.com/>). Since all WAV files had similar size (1.6 MB), we simplified computations by using compressed file size as the predictor.

Expectancy-Violation model. Eerola et al.'s expectancy-based model (EBM; Eerola & North, 2000; Eerola, Himberg, Toivainen, & Louhivuori 2006), later renamed Expectancy-Violation model (EV; Eerola, 2016), is a feature-based

model. Concretely, we used the EV₄ model (Eerola, 2016) with predictors: tonal ambiguity, pitch proximity, entropy of duration distribution, and entropy of pitch-class distribution. This validated instrument is in line with our design, including some of the parameters we manipulated to characterize the Complexity subset, and is thus preferred over other models such as Streich's (2007). As pointed out by Albrecht (2016), Eerola's (2016) study convincingly indicated that just a few low-level parameters could predict a relatively large portion of the variance in judgments of perceived melodic complexity. Eerola's model has been used to assess melodic complexity in several studies, such as Fiveash, McArthur, and Thompson (2018), and, more generally, musical features in Albrecht (2018).

Information Dynamics of Music model. The IDyOM (Pearce, 2005; Pearce, 2018) is a variable-order Markov model (Begleiter, El-Yaniv, & Yona, 2004; Bunton, 1997) that uses a multiple-viewpoint framework (Conklin & Witten, 1995), allowing it to combine models of different representations of the musical surface. IDyOM has been shown to accurately predict Western listeners' pitch expectations in behavioral, physiological, and EEG studies (e.g., Eggermann et al., 2013; Hansen & Pearce, 2014; Omigie, Pearce, & Stewart, 2012; Omigie, Pearce, Williamson, & Stewart, 2013; Pearce, 2005; Pearce, Ruiz, Kapasi, Wiggins, & Bhattacharya, 2010), even better than static rule-based models (e.g., Narmour, 1991; Schellenberg, 1997). It has also been proved to account for expectations of the timing of melodic events (Sauvé, Sayed, Dean, & Pearce, 2018) and harmonic movement (Sears, Pearce, Spitzer, Caplin, & McAdams, 2018; Harrison & Pearce, 2018), and to simulate other psychological processes in music perception, including similarity perception (Pearce & Müllensiefen, 2017), recognition

memory (Agres, Abdallah, & Pearce, 2018), phrase boundary perception (Pearce, Müllensiefen, & Wiggins, 2010), and aspects of emotional experience (Egermann et al., 2013; Gingras et al., 2016; Sauvé et al., 2018). We used the IDyOM in two configurations: first, the short-term model (STM) that learns incrementally on each stimulus independently; second, adding to the STM a long-term model (LTM) trained on a large corpus of Western tonal music (903 folk songs and chorales; datasets 1, 2, and 9 from Table 4.1 in Pearce, 2005, comprising 50,867 notes): the BOTH configuration. This incorporates a learned model of schematic musical syntax, providing a measure of complexity relative to the norms of the Western tonal musical style. Both configurations predict the pitch and onset of every note using a combined representation of melodic pitch interval and tonal scale degree (for pitch), and inter-onset interval ratio (in the case of onset).

To compare our composite computational measure of perceived complexity with the models described above (FLAC, EV₄, and IDyOM in its two configurations), we first conducted four linear mixed-effects models. Participants' ratings were modeled using each motif's complexity estimate produced by FLAC, EV₄, and IDyOM in its two configurations, as the independent variable. The design was similar to the complexity model described above. We compared the results of these models to the results of our MUST_K model using likelihood ratio tests. For statistically significant differences ($p < .050$), lower Bayesian information criterion (BIC) and Akaike information criterion (AIC) indicate a better fit of one model over another.

Results

Computational measures of musical attributes.

Appendix C in the Supplementary Materials collects the values of each of the computational measures and components for each of the 200 stimuli.

Composite nonredundant measures.

Balance. The computational measures in the Balance subset were adequate for PCA (Bartlett's: $\chi^2_{(3)} = 173.822$, $p < .001$; Overall MSA = .75, MSA_{Bisect unbalance} = .86; MSA_{Center of mass offset} = .68; MSA_{Event heterogeneity} = .74; Determinant of the correlation matrix = .025). The PCA with *oblimin* rotation indicated that the three initial Balance measures could be subsumed into a single component explaining 95% of the variance. The three measures contributed with similar high loadings (*bisect unbalance*: .95; *center of mass offset*: .98; *event heterogeneity*: .97). We calculated the component scores for each stimulus (BC1) and regarded these as their Balance scores (Table C1, Appendix C).

Contour. The computational measures in the Contour subset were suitable for PCA (Bartlett's: $\chi^2_{(6)} = 135.974$, $p < .001$); Overall MSA = .70; MSA_{Average absolute interval} = .66; KMO_{Melodic abruptness} = .65; KMO_{Durational abruptness} = .85; KMO_{Rhythmic abruptness} = .64; Determinant of the correlation matrix = .055). The PCA indicated that we should extract two components according to Jolliffe's criterion (eigenvalue PC1 = 2.79; eigenvalue PC2 = 0.87), explaining 91% of the variance. After *oblimin* rotation, CC1 represented 71% of the explained variance and received loadings from *average absolute interval* (.99), *melodic abruptness* (.95), and *durational abruptness* (.81). *Rhythmic abruptness* corresponded to CC2 with a loading of .99.

The component scores for each of the stimuli constituted their Contour (CC1 and CC2) scores (Table C2, Appendix C).

Symmetry. The computational measures in the Symmetry subset were suitable for PCA (Bartlett's: $\chi^2_{(1)} = 92.403$, $p < .001$; Overall MSA = .50; MSA *Total asymmetry* = .50; MSA *Asymmetry index* = .50; Determinant of the correlation matrix = .143). The PCA resulted in a single component with eigenvalue 1.93, explaining 96% of variance, and comprising *total asymmetry* and *asymmetry index* with equal contributions of .98. The component score for each stimulus (SC1) represented its Symmetry score (Table C3, Appendix C).

Complexity. We first checked whether the data set was adequate for PCA. The determinant of the correlation matrix was lower than 10^{-5} , meaning that there was too much redundancy in the data. Due to excessive multicollinearity, we removed variables with high correlations with other variables: *pitch entropy*, *2-tuple pitch entropy*, and *3-tuple pitch entropy*. The remaining computational measures in the Complexity subset were suitable for PCA (Bartlett's: $\chi^2_{(10)} = 246.082$, $p < .001$; Overall MSA = .73; MSA *Event density* = .78; MSA *Average local pitch entropy* = .75; MSA *2-tuple interval entropy* = .71; MSA *3-tuple duration entropy* = .65; MSA *Weighted permutation entropy* = .68; Determinant of the correlation matrix = .005). The PCA indicated that two components should be extracted according to Jolliffe's criterion (eigenvalue $_{PC1} = 3.47$; eigenvalue $_{PC2} = 1.00$), explaining 89% of the variance. After *oblimin* rotation, KC1 comprised *event density* (1.00), *average local pitch entropy* (.96), *2-tuple interval entropy* (.94), and *weighted permutation entropy* (.60). These measures quantified the number of elements and pitch entropies, and accounted for 72% of the explained variance. KC2 corresponded to *3-tuple duration entropy*

(.96). The component scores for each stimulus became their Complexity (KC1 and KC2) scores (Table C4, Appendix C).

Explaining participants' ratings of musical attributes.

The results with the outliers included in the analysis described here are reported in Appendix E in the Supplementary Materials.

Balance. After removing three participants whose ratings were highly influential according to Cook's distances, and rerunning the model, the linear mixed-effects model showed that the component calculated in the PCA reported above (BC1) was a strong predictor of participants' balance ratings ($\beta = 0.925$, $t_{(38.952)} = 7.992$, $p < .001$). The effect of BC1 was medium to large ($d = 0.72$).

Contour. The only participant whose Cook's distances were above the threshold was removed from the model, which was then run again. The new linear mixed-effects model of contour showed that both components resulting from the PCA were strong predictors of participants' ratings of contour (CC1: $\beta = 0.774$, $t_{(41.053)} = 8.474$, $p < .001$; CC2: $\beta = 0.370$, $t_{(48.123)} = 6.813$, $p < .001$). The interaction effect was also significant ($\beta = -0.221$, $t_{(57.200)} = -6.298$, $p < .001$), meaning that the stronger the influence of one component on participants' ratings, the weaker the influence of the other component. CC1 had a medium to large effect ($d = 0.61$), CC2 had a small to medium effect ($d = 0.29$), and the CC1*CC2 interaction had a small effect ($d = 0.17$).

Symmetry. When the highly influential participant had been removed, the linear mixed-effects model revealed that the Symmetry component (SC1) produced by the PCA was a strong predictor of participants' ratings of symmetry ($\beta = 0.380$, $t_{(40.934)} = 5.410$, $p < .001$). The effect of SC1 was small ($d = 0.24$).

Complexity. One participant highly influenced the model, and was therefore removed. The resulting linear mixed-effects model revealed that both components resulting from the PCA were strong predictors of participants' complexity ratings, KC1 ($\beta = 1.183$, $t_{(41.409)} = 30.729$, $p < .001$) and KC2 ($\beta = 0.140$, $t_{(45.394)} = 5.322$, $p < .001$). In addition, a mutually enhancing interaction between components was also significant ($\beta = 0.139$, $t_{(116.995)} = 5.991$, $p < .001$). KC1 had a very large effect ($d = 1.26$), KC2 had a small effect ($d = 0.15$), and so did the KC1*KC2 interaction ($d = 0.15$).

Comparison with existing models of perceived complexity.

The four new linear mixed-effects models showed that other existing models of musical complexity were also good predictors of participants' complexity ratings (Table 3). However, the ANOVA mixed model likelihood ratio tests showed that our model provided a better fit to the data than all but one of the extant complexity models. Although the IDyOM STM provided a better fit to the data than our MUST_K model according to AIC and BIC, the difference was not statistically significant (Table 4).

Table 3. *Linear Mixed-effects Models of Complexity for the Complexity Subset*

Model	Component	β	df	t -value	p -value	d
MUST_K	KC1	1.18	41.41	30.729	< .001	1.26
	KC2	0.14	45.39	5.322	< .001	0.15
	KC1*KC2	0.139	116.995	5.991	< .001	0.15
FLAC		0.999	40.391	39.41	< .001	0.94
EV₄		1.106	41.179	37.81	< .001	1.16
IDyOM (STM)		1.146	40.828	39.71	< .001	1.27
IDyOM (BOTH)		1.074	40.691	37.25	< .001	1.09

Note. The models of perceived complexity compared here are the MUST_K model, FLAC compression size, the Expectancy-Violation model with four predictors (EV₄), and the Information Dynamics of Music model (IDyOM) in the short-term (STM) and BOTH configurations. β refers to the estimated slope, df to the degrees of freedom, and d to the effect size.

Table 4. *ANOVA Mixed Model Likelihood Ratio Tests of Comparisons with the MUST_K Model*

Model	<i>df</i>	AIC	BIC	logLik	$\chi^2_{(9)}$	<i>p</i>
MUST_K	15	4986.5	5069.7	-2478.2		
FLAC	6	5526.3	5559.5	-2757.1	557.76	< .001
EV₄	6	5058.6	5091.9	-2523.3	90.127	< .001
IDyOM (STM)	6	4829.3	4862.5	-2408.6	0	1
IDyOM (BOTH)	6	5214.0	5247.3	-2601.0	245.53	< .001

Note. The models of perceived complexity compared here are the MUST_K model, FLAC compression size, the Expectancy-Violation model with four predictors (EV₄), and the Information Dynamics of Music model (IDyOM) in the short-term (STM) and BOTH configurations. The table shows the degrees of freedom (*df*), the Akaike information criterion (AIC), the Bayesian information criterion (BIC), the log likelihood (logLik), and the *p*-value for each model comparison. The chi-squared value (χ^2) for each particular model involved 9 degrees of freedom for all models compared.

Discussion

This second study focused on the structural features of the 200 musical motifs we created. We had four main goals. The first was to devise a series of computational measures providing objective descriptions of the parameters manipulated in the composition of the motifs. This led us to develop three measures of balance, four measures of contour, three measures of symmetry, and eight measures of complexity. They can be used for diverse purposes in conjunction with our stimulus set or applied to other musical motifs.

The computational measures were designed to capture aspects of the same attribute, so they were bound to include a certain degree of redundancy and multicollinearity. Our second goal was thus to derive nonredundant composite measures for each of the four attributes using principal component analyses

(PCA). The results of the PCA for balance revealed that the three measures loaded highly on a single component (BC1), indicating that, in our Balance subset, the three parameters (distribution of elements/events, climax position, tension) work together to create different degrees of balance and unbalance. The composite of the three measures, calculated as the component score, constitutes each of the musical motifs' Balance score. The PCA for contour revealed two components underlying the computational measures (CC1 and CC2). The three measures of intervallic and melodic abruptness loaded onto one component (CC1), and the measure of rhythmic abruptness loaded onto another (CC2), thus mirroring the two parameters used to compose the motifs in the Contour subset, and provide the musical motifs' Contour scores. The PCA for symmetry subsumed both computational measures into a single component (SC1), in accordance with our manipulation of a single aspect of symmetry: vertical mirror structure. The composite of both measures, calculated as the component score, is each motif's Symmetry score. Finally, the PCA for complexity revealed two components underlying the computational measures (KC1 and KC2): The first was related to the number of elements, and variety and predictability of pitches, whereas the second was related to the variety and predictability of durations, thus reflecting the aspects underlying variations in the complexity of the motifs in the Complexity subset, and constitute the motifs' Complexity scores.

Our third goal was to examine the extent to which the composite measures, or combination thereof, explained participants' ratings in Study 1. The linear mixed-effects models for each attribute showed that the composite measures were strong predictors of perceived balance (BC1), contour (CC1 and

CC2), symmetry (SC1), and complexity (KC1 and KC2). The results also revealed an interaction between the Contour components (CC1 and CC2) meaning that, while both individually serve as predictors of perceived musical smoothness or jaggedness, when one component (e.g., intervallic and melodic abruptness) exerts a higher influence on participants' ratings, the effect of the other (e.g., rhythmic abruptness) becomes smaller. There was also an interaction between the Complexity components (KC1 and KC2). In this case, there was a mutual enhancement: when one component (e.g., number and variety of events) exerts a stronger influence on participants' ratings, so does the other (e.g., number and variety of durations), contributing to musical complexity in complementarily reinforcing ways.

A closer look at the relations between the design and the assessments may help to understand the processes involved in the perception of these attributes in music, enabling comparison with other sensory modalities. The results suggest that our balance measures indeed captured the tensional processes and temporal discourse of the motifs, which in turn seem largely responsible for the perception of musical balance. Likewise, both pitch and rhythm correspondences between the halves of the motif appear equally relevant for the perception of musical symmetry. A different pattern emerged for perceived contour: The results suggest that an enhanced salience of either pitch (CC1) or rhythm (CC2) relations due to a pronounced abruptness reduces the prominence of the other dimension. In contrast, for complexity, the quantity and variety of elements together with pitch-related order or structure (KC1), and rhythm-related order or structure (KC2) reinforce each other in their impact on perceived musical complexity. As the most

salient dimension in Western tonal music, pitch relations define harmony and structure rhythm, which reciprocally modulates pitch relations (Prince, 2011; Prince, Thomson, & Schmuckler 2009).

Two inversely related factors mainly account for perceived visual complexity: quantity and variety of elements, and order or structure (e.g., Gartus & Leder, 2017; Nadal et al., 2010). They also constitute the core of perceived complexity in music, and their interrelations in the temporal and spatial dimensions deserve close attention. The various measures of entropy assessed order and structure in music, inevitably integrating variety of elements and predictability. These factors are interdependent, and the investigation of their relative contributions would require controlling for one while manipulating the other within a common idiom.

Pitch-related entropies naturally correlated with quantity of elements, the best individual predictor also in visual studies: Maximal pitch-related entropies increase with the number of elements (equivalent to *event density*, in our case)—although this relationship saturates at a certain point, as *event density* is restricted by the musical idiom: the variety of sounds is constrained, as the notes are discrete and we established a vocal pitch range. Therefore, even though there is no theoretical boundary for maximal entropy, it is, in practice, limited by the musical style. To discern the particular contributions of pitch-related entropies, controlling for *event density* would be required. In contrast, *duration entropy* (order and structure in time) is always constrained by *event density* (number of elements). The different contributions of pitch and rhythm to perceived musical complexity also respond to the combination of several factors: First, the number of different

rhythmic figures is lower than that of pitches in this particular musical idiom. Second, ratios are better recognized and remembered than absolute values (Pressing, 1999; Trehub, 1985), and pattern transformation techniques are standard compositional techniques (e.g., augmentation, retrogradation), all of which limit the number of combinations appraised as different.

Testing our computational models with other musical stimuli would either strengthen or question the validity of our approach and throw light on the way humans perceive such attributes in music. This was only possible for complexity, because no comparable computational assessments of perceived musical balance, contour, and symmetry, as defined in the stimulus design, are available. The fourth goal of this study was to compare the explanatory performance of our MUST_K model with other approaches to perceived musical complexity. The four extant models we used for comparison proved to be good predictors of participants' ratings. This suggests that they all tap into the same phenomenon. However, according to the model likelihood ratio tests and under the AIC and BIC criteria, they do so to different extents. Our model predicted participants' ratings more accurately than FLAC compression, EV₄, and the BOTH configuration of the IDyOM. The STM configuration, which generates predictions after learning directly from each specific stimulus, provided the best fit to participants' complexity ratings, though not significantly better than the MUST_K model developed here.

The better fit provided by our model might not be surprising, taking into account that it addresses precisely the design features of the musical motifs in the Complexity subset. Nevertheless, it is worth noting some differences between the

parameters included in these models. The superiority over the EV₄ model can be explained by the motifs' common idiom that might have lessened the effect of EV₄'s first component (*tonal ambiguity*), but also by a more comprehensive design and better performance of our measures—e.g., EV₄ considers pitch-class instead of absolute pitch, which ignores the contribution of pitch height across different octaves to perceived complexity. Investigating whether this applies to other musical stimuli would shed light on the factors underlying perceived musical complexity.

The comparisons with the FLAC and IDyOM models are especially noteworthy. A higher predictive capacity over the FLAC general-purpose audio compression algorithm may be due to the encoding of high-level symbolic features that are specific to the musical language in our model compared with the raw audio input for FLAC (sampled at 44,100 Hz with a bit depth of 16). Elucidating whether our model's superiority generalizes to other musical stimuli would shed some light on the processing of musical complexity: If our model surpassed FLAC's prediction power with other music beyond the present stimulus set, the perception of musical complexity would be driven by the combination of irreducible, basic musical features. If this were not the case, the implication would be that musical complexity is holistically appraised using general-purpose perceptual processes.

Regarding the IDyOM models, the fact that the simulation of participants' musical background worsens the short-term model may seem striking. However, it is perhaps not surprising that the BOTH model does less well than the STM and the MUST_K model, because the stimuli are stylistically coherent, and complexity

does not vary as a function of distance from Western tonal stylistic norms. This means that the BOTH configuration addresses the issue of context or previous experience not as a framework in which to discriminate degrees of complexity, but as a form of averaged reference from which to detect deviations. On the other hand, the MUST_K model employs features crafted with knowledge of the stimulus design and was fitted to the perceptual responses to the stimuli, whereas the IDyOM complexity measures were generated entirely without prior knowledge of either the stimulus set itself or the perceptual complexity ratings for these stimuli. However, the STM learns directly from the stimulus, and thus the adaptation to the stimulus set may be similar. But more importantly, the MUST_K model is based on low-level musical parameters, less computationally demanding than the STM, and thus more parsimonious. Therefore, the lack of significant differences in predictive power between these two models supports the validity of our approach and suggests that the processing of musical complexity relies on isolable basic features as those captured by the MUST_K model. Further research with other stimuli will elucidate whether the present results generalize to the perceived complexity of any music.

General Discussion

Choosing among alternative options and courses of action is one of the most basic functions of cognition. Understanding cognition, therefore, requires understanding the processes involved in the valuation and comparison of alternatives. There are several reasons why music constitutes a rich domain for studying general mechanisms of valuation: Music provides a rich and virtually unlimited set of materials and is highly valued among people. But it also affords

an investigation of the interaction between domain-specific and domain-general processes in valuation. The overarching goal of the research presented in this paper was to stimulate research on modality-general attributes and domain-general processes in the appreciation of music. We set out (1) to create a set of musical stimuli suitable for studying the role of modality-general attributes in music, (2) to assess the stimuli behaviorally and computationally, (3) to analyze how both kinds of assessments relate to each other, the stimulus design features, and other available measures, and (4) to make the MUST set and computational measures in the form of a MATLAB toolbox freely available to other researchers.

The design of the four subsets responds to a modality-general characterization of balance, contour, symmetry, and complexity: We distilled the essence of three sets of visual stimuli (Wilson & Chatterjee, 2005, for balance; Bertamini et al., 2016, for contour; and Jacobsen & Höfel, 2002, for symmetry and complexity) and formulated analogous musical definitions for each attribute. We restricted the design to a common idiom that makes the motifs comparable to the emulated visual stimuli and allows contrasting the target attributes across different musical examples.

Our stimuli and computational measures contribute to the investigation of perceived musical balance, contour, and symmetry in music, and further explore perceived musical complexity. Whereas the existing literature on musical complexity is comparable to that in the visual domain, a small number of studies address musical symmetry (e.g., Balch, 1981; Bianchi et al., 2017; Krumhansl et al., 1987; Mongoven & Carbon, 2017), while others investigate musical contour (e.g., Gerardi & Gerken, 1995; Schmuckler, 2015; Thorpe, 1986; Trehub et al.,

1984). To the best of our knowledge, our research pioneers the study of musical balance as conceived here, and our modality-general characterization of these four attributes within a coherent set and toolbox is a unique contribution.

The MUST set combines ecological validity and experimental control, a delicate and desirable balance between two core virtues of any set of stimuli. The results demonstrated that the set is sensitive to nonmusicians' abilities to detect degrees of musical balance, complexity, contour, and symmetry (cf., Petrović, et al., 2017), accurately captured by the computational measures: Participants' consistent judgments matched the stimulus design and were largely explained by our composite models. Furthermore, the comparisons with extant models of musical complexity support ours as an outstanding approach. The coherence between design and assessments strengthens the value of the set and the computational measures as reliable open resources for research. First, its virtues make the set highly useful in empirical aesthetics and other fields, especially in its abridged form and when the interest is musical–visual correspondence. Second, the measures contribute new tools to music information research because they may easily be applied to other stimuli. Ultimately, investigating the relations between the stimulus design, their behavioral appraisal, and the computational measures may contribute to further understanding of musical and psychological processes.

The MUST stimuli and computational measures may be useful in multiple settings and fields, together or separately: First, the subsets may be used together addressing several attributes or individually focusing on one of them, and the motifs can be assessed in other ways. Indeed, the design of other assessments is

feasible and desirable, especially regarding the less studied attributes. Second, while the measures perfectly complement the stimuli, their general character and reliable performance in predicting participants' judgments make them suitable for other purposes and musical stimuli as well, even if small adaptations were needed. Monophonic melodies would be particularly appropriate, especially if short, for which no specific adjustment would be required. However, testing them with longer, more varied, and naturalistic musical stimuli would be of great interest in assessing how the measures and fitted models generalize as models of music perception. To facilitate the use of the methods and materials presented here by other researchers, we have made the full and abridged stimulus set, together with the open-source package of functions as a toolbox for MATLAB, freely available for use by the scientific community at osf.io/bfxz7. The detailed description and formulation of the measures constitute Appendix B, and the values for each stimulus in each of the corresponding measures and components constitute Appendix C of the Supplementary Materials.

Authors' Contributions

AC created the stimuli and wrote the manuscript; AC and MV designed the computational measures; MV formalized, implemented, and wrote the measures; AC and MN designed the research, discussed the stimuli, and analyzed the data; AC, GC, GA, and MN contributed to the behavioral assessment; AC, MV, MP, and MN compared and discussed the measures, and revised the manuscript. All authors reported no conflicts of interest and approved the manuscript.

Funding Information

The project leading to these results has received funding from —La Caixa Foundation (ID 100010434) under agreements LCF/BQ/ES17/11600021 and LCF/BQ/DE17/11600022, and from the Spanish Ministerio de Economía, Industria y Competitividad with grant PSI2016- 77327-P.

References

- Aguinis, H., Gottfredson, R. K., & Joo, H. (2013). Best-Practice Recommendations for Defining, Identifying, and Handling Outliers. *Organizational Research Methods*, 16(2), 270–301. doi:10.1177/1094428112470848
- Agres, K., Abdallah, S., & Pearce, M. (2018). Information-Theoretic Properties of Auditory Sequences Dynamically Influence Expectation and Memory. *Cognitive science*, 42(1), 43–76. doi:10.1111/cogs.12477
- Albrecht, J. (2016). Modeling Musical Complexity: Commentary on Eerola (2016). *Empirical Musicology Review*, 11(1), 20. doi:10.18061/emr.v11i1.5197
- Albrecht, J. D. (2018). Expressive Meaning and the Empirical Analysis of Musical Gesture. *Music Theory Online*, 24(3). doi:10.30535/mt0.24.3.1
- Baayen, R. H., Davidson, D. J., & Bates, D. M. (2008). Mixed-effects modeling with crossed random effects for subjects and items. *Journal of memory and language*, 59(4), 390–412. doi:10.1016/j.jml.2007.12.005
- Balch, W. R. (1981). The role of symmetry in the good continuation ratings of two-part tonal melodies. *Perception & Psychophysics*, 29(1), 47–55. doi:10.3758/bf03198839
- Barr, D. J., Levy, R., Scheepers, C., & Tily, H. J. (2013). Random effects structure for confirmatory hypothesis testing: Keep it maximal. *Journal of memory and language*, 68(3), 255–278. doi:10.1016/j.jml.2012.11.001
- Bates, D., Maechler, M., Bolker, B., & Walker, S. (2015). Fitting Linear Mixed-Effects Models Using lme4. *Journal of Statistical Software*, 67(1), 1–48. doi:10.18637/jss.v067.i01.
- Begleiter, R., El-Yaniv, R., & Yona, G. (2004). On prediction using variable order Markov models. *Journal of Artificial Intelligence Research*, 22, 385–421. doi:10.1613/jair.1491
- Berridge, K. C., & Kringelbach, M. L. (2013). Neuroscience of affect: brain mechanisms of pleasure and displeasure. *Current Opinion in Neurobiology*, 23(3), 294–303. doi:10.1016/j.conb.2013.01.017
- Bertamini, M., Palumbo, L., Gheorghes, T. N., & Galatsidas, M. (2016). Do observers like curvature or do they dislike angularity?. *British Journal of Psychology*, 107(1), 154–178. doi:10.1111/bjop.12132
- Bianchi, I., Burro, R., Pezzola, R., & Savardi, U. (2017). Matching Visual and Acoustic Mirror Forms. *Symmetry*, 9(3), 39. doi:10.3390/sym9030039
- Brattico, E., & Pearce, M. T. (2013). The neuroaesthetics of music. *Psychology of Aesthetics, Creativity, and the Arts*, 7, 48–61. doi:10.1037/a0031624
- Brieber, D., Nadal, M., Leder, H., & Rosenberg, R. (2014). Art in time and space: context modulates the relation between art experience and viewing time. *PloS ONE*, 9(6), e99019. doi:10.1371/journal.pone.0099019
- Bunton, S. (1997). Semantically motivated improvements for PPM variants. *The Computer Journal*, 40(2/3), 76–93. doi:10.1093/comjnl/40.2_and_3.76
- Caplin, W. E., Hepokoski, J., & Webster, J. (2010). *Musical Form, Forms & Formenlehre*, Leuven University Press. doi:10.2307/j.ctt9qf01v
- Cattaneo, Z., Lega, C., Ferrari, C., Vecchi, T., Cela-Conde, C. J., Silvanto, J., & Nadal, M. (2015). The role of the lateral occipital cortex in aesthetic appreciation of representational and

- abstract paintings: A TMS study. *Brain and Cognition*, 95, 44–53. doi:10.1016/j.bandc.2015.01.008
- Che, J., Sun, X., Gallardo, V., & Nadal, M. (2018). Cross-cultural empirical aesthetics. The Arts and The Brain - Psychology and Physiology Beyond Pleasure, *Progress in Brain Research*, 237, 77–103. doi:10.1016/bs.pbr.2018.03.002
- Coalson, J. (2008). Flac-free lossless audio codec. Retrieved from <http://flac.sourceforge.net> (1/11/2018)
- Conklin, D., & Witten, I. H. (1995). Multiple viewpoint systems for music prediction. *Journal of New Music Research*, 24(1), 51–73. doi:10.1080/09298219508570672
- Cook, N. (1987). Musical form and the listener. *The Journal of aesthetics and art criticism*, 46(1), 23–29. doi:10.2307/431305
- Cook, R. D. (1979). Influential observations in linear regression. *Journal of the American Statistical Association*, 74(365), 169–174.
- Corradi, G., Chuquichambi, E. G., Barrada, J. R., Clemente, A., & Nadal, M. (2020). A new conception of aesthetic sensitivity. *British Journal of Psychology*. doi:10.1111/bjop.12427
- Cross, I. (2006). Music, Cognition, Culture, and Evolution. *Annals of the New York Academy of Sciences*, 930(1), 28–42. doi:10.1111/j.1749-6632.2001.tb05723.x
- De Lange, F. P., Heilbron, M., & Kok, P. (2018). How Do Expectations Shape Perception? *Trends in Cognitive Sciences*, 22(9), 764–779. doi:10.1016/j.tics.2018.06.002
- Dissanayake, E. (2008). If music is the food of love, what about survival and reproductive success? *Musicae Scientiae*, 12(1_suppl), 169–195. doi:10.1177/1029864908012001081
- Edmonston, W. E. Jr. (1969). Familiarity and Musical Training in the Esthetic Evaluation of Music. *The Journal of Social Psychology*, 79(1), 109–111. doi:10.1080/00224545.1969.9922393
- Eerola, T. (2016). Expectancy-violation and information-theoretic models of melodic complexity. *Empirical Musicology Review*, 11(1), 2–17. doi:10.18061/emr.v11i1.4836
- Eerola, T., Himberg, T., Toiviainen, P., & Louhivuori, J. (2006). Perceived complexity of Western and African folk melodies by Western and African listeners. *Psychology of Music*, 34(3), 337–371. doi:10.1177/0305735606064842
- Eerola, T., & North, A. C. (2000, August). Expectancy-based model of melodic complexity. In *Proceedings of the Sixth International Conference on Music Perception and Cognition*. Keele, Staffordshire, UK: Department of Psychology. CD-ROM.
- Egermann, H., Pearce, M. T., Wiggins, G. A., & McAdams, S. (2013). Probabilistic models of expectation violation predict psychophysiological emotional responses to live concert music. *Cognitive, Affective, & Behavioral Neuroscience*, 13(3), 533–553. doi:10.3758/s13415-013-0161-y
- Fiveash, A., McArthur, G., & Thompson, W. F. (2018). Syntactic and non-syntactic sources of interference by music on language processing. *Scientific Reports*, 8(1). doi:10.1038/s41598-018-36076-x
- Forsythe, A., Mulhern, G., & Sawey, M. (2008). Confounds in pictorial sets: The role of complexity and familiarity in basic-level picture processing. *Behavior Research Methods*, 40(1), 116–129. doi:10.3758/brm.40.1.116

- Forsythe, A., Nadal, M., Sheehy, N., Cela-Conde, C. J., & Sawey, M. (2011). Predicting beauty: Fractal dimension and visual complexity in art. *British Journal of Psychology*, 102, 49–70. doi:10.1348/000712610x498958
- Gartus, A., & Leder, H. (2013). The Small Step toward Asymmetry: Aesthetic Judgment of Broken Symmetries. *I-Perception*, 4(5), 361–364. doi:10.1068/i0588sas
- Gartus, A., & Leder, H. (2017). Predicting perceived visual complexity of abstract patterns using computational measures: The influence of mirror symmetry on complexity perception. *PloS ONE*, 12(11), e0185276. doi:10.1371/journal.pone.0185276
- Gerardi, G. M., & Gerken, L. (1995). The Development of Affective Responses to Modality and Melodic Contour. *Music Perception: An Interdisciplinary Journal*, 12(3), 279–290. doi:10.2307/40286184
- Gingras, B., Pearce, M. T., Goodchild, M., Dean, R. T., Wiggins, G., & McAdams, S. (2016). Linking melodic expectation to expressive performance timing and perceived musical tension. *Journal of Experimental Psychology: Human Perception and Performance*, 42(4), 594–609.
- Gómez-Puerto, G., Munar, E., & Nadal, M. (2015). Preference for curvature: A historical and conceptual framework. *Frontiers in Human Neuroscience*, 9, 712. doi:10.3389/fnhum.2015.00712
- Grey, T. S. (1988). Wagner, the Overture, and the Aesthetics of Musical Form. *19th-Century Music*, 12(1), 3–22. doi:10.1525/nem.1988.12.1.02a00010
- Hansen, N. C., & Pearce, M. T. (2014). Predictive uncertainty in auditory sequence processing. *Frontiers in Psychology*, 5, 1052. doi:10.3389/fpsyg.2014.01052
- Harrison, P., & Pearce, M. T. (2018). An energy-based generative sequence model for testing sensory theories of Western harmony. *arXiv preprint arXiv:1807.00790*.
- Heyduk, R. G. (1975). Rated preference for musical compositions as it relates to complexity and exposure frequency. *Perception & Psychophysics*, 17(1), 84–90.
- Hox, J. J., Moerbeek, M., & van de Schoot, R. (2010). *Multilevel analysis: Techniques and applications*. Routledge.
- Huron, D. (2003). Is music an evolutionary adaptation? In I. Peretz & R. Zatorre (Eds.), *The cognitive neuroscience of music*, 57–75. doi:10.1093/acprof:oso/9780198525202.003.0005
- Jacobsen, T., & Höfel, L. E. A. (2002). Aesthetic judgments of novel graphic patterns: analyses of individual judgments. *Perceptual and Motor Skills*, 95(3), 755–766. doi:10.2466/pms.2002.95.3.755
- Jakesch, M., & Leder, H. (2015). The qualitative side of complexity: Testing effects of ambiguity on complexity judgments. *Psychology of Aesthetics, Creativity, and the Arts*, 9, 200–205. doi:10.1037/a0039350
- Jolliffe, I. T. (1972). Discarding Variables in a Principal Component Analysis. I: Artificial Data. *Applied Statistics*, 21(2), 160. doi:10.2307/2346488
- Judd, C. M., Westfall, J., & Kenny, D. A. (2017). Experiments with more than one random factor: designs, analytic models, and statistical power. *Annual Review of Psychology*, 68(1), 601–625. doi:10.1146/annurev-psych-122414-033702

- Juslin, P. N. (2013). From everyday emotions to aesthetic emotions: Towards a unified theory of musical emotions. *Physics of Life Reviews*, 10(3), 235–266. doi:10.1016/j.plrev.2013.05.008
- Koelsch, S., Vuust, P., & Friston, K. (2018). Predictive Processes and the Peculiar Case of Music. *Trends in Cognitive Sciences*, 23(1), 63–77. doi:10.1016/j.tics.2018.10.006
- Kringelbach, M. L., & Berridge, K. C. (2009). Towards a functional neuroanatomy of pleasure and happiness. *Trends in Cognitive Sciences*, 13(11), 479–487. doi:10.1016/j.tics.2009.08.006
- Krumhansl, C. L., Sandell, G. J., & Sergeant, D. C. (1987). The Perception of Tone Hierarchies and Mirror Forms in Twelve-Tone Serial Music. *Music Perception: An Interdisciplinary Journal*, 5(1), 31–77. doi:10.2307/40285385
- Kuznetsova, A., Brockho, P. B., & Christensen, R. H. B. (2012). lmerTest: Tests for random and fixed effects for linear mixed effect models (lmer objects of lme4 package). Retrieved from <http://www.cran.r-project.org/package=lmerTest/> (1/11/2018)
- Leichtentritt, H. (1911). *Musikalische Formenlehre* (Vol. 8). Breitkopf & Härtel.
- Levy, D. J., & Glimcher, P. W. (2012). The root of all value: a neural common currency for choice. *Current Opinion in Neurobiology*, 22(6), 1027–1038. doi:10.1016/j.conb.2012.06.001
- Locher, P., Gray, S., & Nodine, C. (1996). The structural framework of pictorial balance. *Perception*, 25, 1419–1436. doi:10.1068/p251419
- Machado, P., Romero, J., Nadal, M., Santos, A., Correia, J., & Carballal, A. (2015). Computerized measures of visual complexity. *Acta Psychologica*, 160, 43–57. doi:10.1016/j.actpsy.2015.06.005
- Madison, G., & Schiölde, G. (2017). Repeated Listening Increases the Liking for Music Regardless of Its Complexity: Implications for the Appreciation and Aesthetics of Music. *Frontiers in Human Neuroscience*, 11, 147. doi:10.3389/fnhins.2017.00147
- Mallik, A., Chandra, M. L., & Levitin, D. J. (2017). Anhedonia to music and mu-opioids: Evidence from the administration of naltrexone. *Scientific Reports*, 7, 41952. doi:10.1038/srep41952
- Margulis, E. H. (2016). Toward A Better Understanding of Perceived Complexity in Music: A Commentary on Eerola (2016). *Empirical Musicology Review*, 11(1), 18. doi:10.18061/emr.v11i1.5275
- Marin, M. M., Lampatz, A., Wandl, M., & Leder, H. (2016). Berlyne revisited: evidence for the multifaceted nature of hedonic tone in the appreciation of paintings and music. *Frontiers in Human Neuroscience*, 10, 536. doi:10.3389/fnhum.2016.00536
- Marin, M. M., & Leder, H. (2013). Examining complexity across domains: relating subjective and objective measures of affective environmental scenes, paintings and music. *PLoS ONE*, 8(8), e72412. doi:10.1371/journal.pone.0072412
- Mathôt, S., Schreij, D., & Theeuwes, J. (2012). OpenSesame: An open-source, graphical experiment builder for the social sciences. *Behavior Research Methods*, 44(2), 314–324. doi:10.3758/s13428-011-0168-7
- Mongoven, C., & Carbon, C. C. (2017). Acoustic Gestalt: On the perceptibility of melodic symmetry. *Musicae Scientiae*, 21(1), 41–59. doi:10.1177/1029864916637116

- Munar, E., Gómez-Puerto, G., Call, J., & Nadal, M. (2015). Common Visual Preference for Curved Contours in Humans and Great Apes. *PLoS One*, *10*(11), e0141106. doi:10.1371/journal.pone.0141106
- Müllensiefen, D., Gingras, B., Musil, J., & Stewart, L. (2014). The musicality of non-musicians: an index for assessing musical sophistication in the general population. *PloS ONE*, *9*(2), e89642. doi:10.1371/journal.pone.0089642
- Nadal, M., Munar, E., Marty, G., & Cela-Conde, C. J. (2010). Visual complexity and beauty appreciation: Explaining the divergence of results. *Empirical Studies of the Arts*, *28*(2), 173–191. doi:10.2190/em.28.2.d
- Narmour, E. (1991). The top-down and bottom-up systems of musical implication: Building on Meyer's theory of emotional syntax. *Music Perception: An Interdisciplinary Journal*, *9*(1), 1–26. doi:10.2307/40286156
- Nieminen, S., Istók, E., Brattico, E., Tervaniemi, M., & Huotilainen, M. (2011). The development of aesthetic responses to music and their underlying neural and psychological mechanisms. *Cortex*, *47*(9), 1138–1146. doi:10.1016/j.cortex.2011.05.008
- Nieuwenhuis, R., te Grotenhuis, H. F., & Pelzer, B. J. (2012). Influence. ME: tools for detecting influential data in mixed-effects models. *The R-Journal*, *4*(2), 38–47. doi:10.31235/osf.io/a5w4u
- Omigie, D., Pearce, M. T., & Stewart, L. (2012). Tracking of pitch probabilities in congenital amusia. *Neuropsychologia*, *50*(7), 1483–1493. doi:10.1016/j.neuropsychologia.2012.02.034
- Omigie, D., Pearce, M. T., Williamson, V. J., & Stewart, L. (2013). Electrophysiological correlates of melodic processing in congenital amusia. *Neuropsychologia*, *51*(9), 1749–1762. doi:10.1016/j.neuropsychologia.2013.05.010
- Palumbo, L., & Bertamini, M. (2016). The curvature effect: A comparison between preference tasks. *Empirical Studies of the Arts*, *34*, 35–52. doi:10.1177/0276237415621185
- Payne, E. (1980). Towards an Understanding of Music Appreciation. *Psychology of Music*, *8*(2), 31–41. doi:10.1177/030573568082004
- Pearce, M. T. (2005). *The construction and evaluation of statistical models of melodic structure in music perception and composition*. Doctoral dissertation, City University London.
- Pearce, M. T. (2018). Statistical learning and probabilistic prediction in music cognition: mechanisms of stylistic enculturation. *Annals of the New York Academy of Sciences*, *1423*(1), 378–395. doi:10.1111/nyas.13654
- Pearce, M., & Müllensiefen, D. (2017). Compression-based modelling of musical similarity perception. *Journal of New Music Research*, *46*(2), 135–155. doi:10.1080/09298215.2017.1305419
- Pearce, M. T., Müllensiefen, D., & Wiggins, G. A. (2010). The role of expectation and probabilistic learning in auditory boundary perception: A model comparison. *Perception*, *39*(10), 1367–1391. doi:10.1068/p6507
- Pearce, M. T., Ruiz, M. H., Kapasi, S., Wiggins, G. A., & Bhattacharya, J. (2010). Unsupervised statistical learning underpins computational, behavioural, and neural manifestations of musical expectation. *NeuroImage*, *50*(1), 302–313. doi:10.1016/j.neuroimage.2009.12.019

- Pereira, C. S., Teixeira, J., Figueiredo, P., Xavier, J., Castro, S. L., & Brattico, E. (2011). Music and Emotions in the Brain: Familiarity Matters. *PLoS ONE*, 6(11), e27241. doi:10.1371/journal.pone.0027241
- Petrović, M., Ačić, G., & Milanković, V. (2017). Sound of picture vs. picture of sound: musical palindrome. *New Sound: International Magazine for Music*, 50(2), 217–228.
- Pressing, J. (1999). Cognitive complexity and the structure of musical patterns. In *Proceedings of the 4th Conference of the Australasian Cognitive Science Society*.
- Prince, J. B. (2011). The integration of stimulus dimensions in the perception of music. *Quarterly Journal of Experimental Psychology*, 64, 2125–2152. doi:10.1080/17470218.2011.573080
- Prince, J. B., Thompson, W. F., & Schmuckler, M. A. (2009). Pitch and time, tonality and meter: How do musical dimensions combine? *Journal of Experimental Psychology: Human Perception and Performance*, 35, 1598–1617. doi:10.1037/a0016456
- Purwins, H., Grachten, M., Herrera, P., Hazan, A., Marxer, R., & Serra, X. (2008). Computational models of music perception and cognition II: Domain-specific music processing. *Physics of Life Reviews*, 5(3), 169–182. doi:10.1016/j.plrev.2008.03.005
- R Core Team (2018). R: A language and environment for statistical computing. Vienna, Austria: R Foundation for Statistical Computing. Retrieved from <http://www.R-project.org> (1/11/2018)
- Revelle, W. (2018) psych: Procedures for Personality and Psychological Research, Northwestern University, Evanston, Illinois, USA, <https://CRAN.R-project.org/package=psych> Version = 1.8.12.
- Robinson, T. (1994). SHORTEN: Simple lossless and near-lossless waveform compression.
- Rohrmeier, M., Zuidema, W., Wiggins, G. A., & Scharff, C. (2015). Principles of structure building in music, language and animal song. *Philosophical Transactions of the Royal Society B: Biological Sciences*, 370(1664), 20140097–20140097. doi:10.1098/rstb.2014.0097
- Salimpoor, V. N., & Zatorre, R. J. (2013). Neural interactions that give rise to musical pleasure. *Psychology of Aesthetics, Creativity, and the Arts*, 7, 62–75. doi:10.1037/a0031819
- Savage, P. E., Brown, S., Sakai, E., & Currie, T. E. (2015). Statistical universals reveal the structures and functions of human music. *Proceedings of the National Academy of Sciences, USA*, 112, 8987–8992. doi:10.1073/pnas.1414495112
- Sauvé, S. A., Sayed, A., Dean, R. T., & Pearce, M. T. (2018). Effects of pitch and timing expectancy on musical emotion. *Psychomusicology: Music, Mind, and Brain*, 28(1), 17–39. doi:10.1037/pmu0000203
- Schaal, N. K., Banissy, M. J., & Lange, K. (2015). The rhythm span task: comparing memory capacity for musical rhythms in musicians and non-musicians. *Journal of New Music Research*, 44(1), 3–10. doi:10.1080/09298215.2014.937724
- Schellenberg, E. G. (1997). Simplifying the implication-realization model of melodic expectancy. *Music Perception: An Interdisciplinary Journal*, 14(3), 295–318. doi:10.2307/40285723
- Schoenberg, A. (1967). *Fundamentals of musical composition*. Stein, L., & Strang, G., eds. London: Faber & Faber.
- Sears, D. R., Pearce, M. T., Spitzer, J., Caplin, W. E., & McAdams, S. (2018). Expectations for tonal cadences: Sensory and cognitive priming effects. *Quarterly Journal of Experimental Psychology*, 174702181881447. doi:10.1177/1747021818814472

- Schmuckler, M. A. (2015). Tonality and Contour in Melodic Processing. *Oxford Handbooks Online*. doi:10.1093/oxfordhb/9780198722946.013.14
- Shepard, R. N. (1982). Structural representations of musical pitch. In D. Deutsch (Ed.), *Psychology of Music*, 343–390. doi:10.1016/b978-0-12-213562-0.50015-2
- Shmulevich, I., & Povel, D. J. (2000). Measures of temporal pattern complexity. *Journal of New Music Research*, 29(1), 61–69. doi:10.1076/0929-8215(200003)29:01;1-p;ft061
- Shrout, P. E., & Fleiss, J. L. (1979). Intraclass correlations: uses in assessing rater reliability. *Psychological Bulletin*, 86(2), 420–428. doi:10.1037/0033-2909.86.2.420
- Silvia, P. J. (2007). An introduction to multilevel modeling for research on the psychology of art and creativity. *Empirical Studies of the Arts*, 25(1), 1–20. doi:10.2190/6780-361t-3j83-0411
- Snijders, T. A. B., and Bosker, R. J. (2012). *Multilevel analysis. An introduction to basic and advanced multilevel modeling* (2nd ed.). London: SAGE Publications.
- Snyder, B., & Snyder, R. (2000). *Music and memory: An introduction*. MIT press.
- Steck, L., & Machotka, P. (1975). Preference for musical complexity: Effects of context. *Journal of Experimental Psychology: Human Perception and Performance*, 1(2), 170–174. doi:10.1037/0096-1523.1.2.170
- Streich, S. (2007). *Music complexity: A multi-faceted description of audio content*. Doctoral dissertation, University of Pompeu Fabra, Barcelona.
- Thoma, M. V., Ryf, S., Mohiyeddini, C., Ehlert, U., & Nater, U. M. (2012). Emotion regulation through listening to music in everyday situations. *Cognition and Emotion*, 26, 550–560. doi:10.1080/02699931.2011.595390
- Thorpe, L. A. (1986). Perceptual constancy for melodic contour. *Infant Behavior and Development*, 9, 379. doi:10.1016/s0163-6383(86)80385-x
- Tinio, P. P. L., & Leder, H. (2009). Just how stable are stable aesthetic features? Symmetry, complexity, and the jaws of massive familiarization. *Acta Psychologica*, 130, 241–250. doi:10.1016/j.actpsy.2009.01.001
- Trainor, L. J., & Unrau, A. (2011). Development of Pitch and Music Perception. *Springer Handbook of Auditory Research*, 223–254. doi:10.1007/978-1-4614-1421-6_8
- Trehub, S. E. (1985). Auditory Pattern Perception in Infancy. *Auditory Development in Infancy*, 183–195. doi:10.1007/978-1-4757-9340-6_10
- Trehub, S. E., Bull, D., & Thorpe, L. A. (1984). Infants' Perception of Melodies: The Role of Melodic Contour. *Child Development*, 55(3), 821. doi:10.2307/1130133
- Trehub, S. E., & Hannon, E. E. (2006). Infant music perception: Domain-general or domain-specific mechanisms? *Cognition*, 100(1), 73–99. doi:10.1016/j.cognition.2005.11.006
- Van den Bosch, I., Salimpoor, V. N., & Zatorre, R. J. (2013). Familiarity mediates the relationship between emotional arousal and pleasure during music listening. *Frontiers in Human Neuroscience*, 7. doi:10.3389/fnhum.2013.00534
- Van Geert, E., & Wagemans, J. (2019). Order, complexity, and aesthetic appreciation. *Psychology of Aesthetics, Creativity, and the Arts*. doi:10.1037/aca0000224
- Vartanian, O., Navarrete, G., Chatterjee, A., Fich, L. B., Leder, H., Modroño, C., ... Nadal, M. (2019). Preference for curvilinear contour in interior architectural spaces: Evidence from experts and nonexperts. *Psychology of Aesthetics, Creativity, and the Arts*, 13(1), 110–116. doi:10.1037/aca0000150

- Wilson, A., & Chatterjee, A. (2005). The assessment of preference for balance: Introducing a new test. *Empirical Studies of the Arts*, 23(2), 165–180. doi:10.2190/b1lr-mvf3-f36x-xr64
- Winner, E., Rosenblatt, E., Windmueller, G., Davidson, L., & Gardner, H. (1986). Children's perception of 'aesthetic' properties of the arts: Domain-specific or pan-artistic?. *British Journal of Developmental Psychology*, 4(2), 149-160. doi:[10.1111/j.2044-835x.1986.tb01006.x](https://doi.org/10.1111/j.2044-835x.1986.tb01006.x)

Table 1. *Summary of Parameters used to Design the Musical Stimuli in each Subset*

Attribute	Parameter	Feature	
		<i>Balanced</i>	<i>Unbalanced</i>
Balance	Distribution of elements/ events	Regular	Irregular
	Climax position	Centered	Skewed
	Tension	Progressive	Unprepared
Contour		<i>Smooth</i>	<i>Jagged</i>
	Intervals	Only small (≤ 4 ths)	Large (> 4 ths) & small
	Durations	Progressive, small changes	Sudden, large changes
Symmetry		Symmetric	Asymmetric
	Vertical mirror structure	Yes	No
Complexity		<i>Simpler</i>	<i>More complex</i>
	Number of elements/events	Few	Many
	Variety of elements/events	Low	High
	Predictability	High	Low

Table 2. *Computational Measures of the Parameters Used to Compose Musical Motifs Varying in Balance, Contour, Symmetry, and Complexity*

Attribute	Parameter	Computational measure
Balance	Distribution of elements/events	<i>Bisect unbalance</i> : Equilibrium between the two halves of a stimulus
	Climax position	<i>Center of mass offset</i> : Distance between center of mass and geometric center
	Tension	<i>Event heterogeneity</i> : Heterogeneity in the temporal distribution of events
Contour	Intervals	<i>Average absolute interval</i> : Average absolute pitch interval size
		<i>Melodic abruptness</i> : Average interval size of changes of direction per note
		<i>Durational abruptness</i> : Proportion of the stimulus with changes of direction
	Durations	<i>Rhythmic abruptness</i> : Average ratio of consecutive durations
Symmetry	Vertical mirror structure	<i>Total asymmetry</i> : Direct–retrograde accumulated pitch difference <i>Asymmetry index</i> : Proportion of the stimulus with asymmetries
	Number of elements/events	<i>Event density</i> : Number of note events per time unit
Complexity	Variety of elements/events Predictability	<i>Average local pitch entropy</i> : Average pitch entropy of .25-s sliding windows
		<i>Pitch entropy</i> : Entropy of pitch distribution
		<i>2-tuple pitch entropy</i> : Entropy of 2-tuple pitch distribution
		<i>3-tuple pitch entropy</i> : Entropy of 3-tuple pitch distribution
		<i>2-tuple interval entropy</i> : Entropy of 2-tuple interval distribution
		<i>3-tuple duration entropy</i> : Entropy of 3-tuple duration distribution
		<i>Weighted permutation entropy</i> : Permutation entropy considering the <i>SD</i> of the pitch distribution of each 3-note sequence

Table 3. *Linear Mixed-effects Models of Complexity for the Complexity Subset*

Model	Component	β	df	t -value	p -value	d
MUST_K	KC1	1.18	41.41	30.729	< .001	1.26
	KC2	0.14	45.39	5.322	< .001	0.15
	KC1*KC2	0.139	116.995	5.991	< .001	0.15
FLAC		0.999	40.391	39.41	< .001	0.94
EV₄		1.106	41.179	37.81	< .001	1.16
IDyOM (STM)		1.146	40.828	39.71	< .001	1.27
IDyOM (BOTH)		1.074	40.691	37.25	< .001	1.09

Note. The models of perceived complexity here compared are the MUST_K model, FLAC compression size, the Expectancy-Violation model with four predictors (EV₄), and the Information Dynamics of Music model (IDyOM) in the short-term (STM) and BOTH configurations. β refers to the estimated slope, df to the degrees of freedom, and d to the effect size.

Table 4. *ANOVA Mixed Model Likelihood Ratio Tests of Comparisons with the MUST_K Model*

Model	<i>df</i>	AIC	BIC	logLik	$\chi^2_{(9)}$	<i>p</i>
MUST_K	15	4986.5	5069.7	-2478.2		
FLAC	6	5526.3	5559.5	-2757.1	557.76	< .001
EV₄	6	5058.6	5091.9	-2523.3	90.127	< .001
IDyOM (STM)	6	4829.3	4862.5	-2408.6	0	1
IDyOM (BOTH)	6	5214.0	5247.3	-2601.0	245.53	< .001

Note. The models of perceived complexity here compared are the MUST_K model, FLAC compression size, the Expectancy-Violation model with four predictors (EV₄), and the Information Dynamics of Music model (IDyOM) in the short-term (STM) and BOTH configurations. The table informs on the degrees of freedom (*df*), the Akaike Information Criterion (AIC), the Bayesian Information Criterion (BIC), the log likelihood (logLik), and the *p*-value for each model comparison. The chi-squared value (χ^2) for each particular model involved 9 degrees of freedom for all compared models.

Figure 1

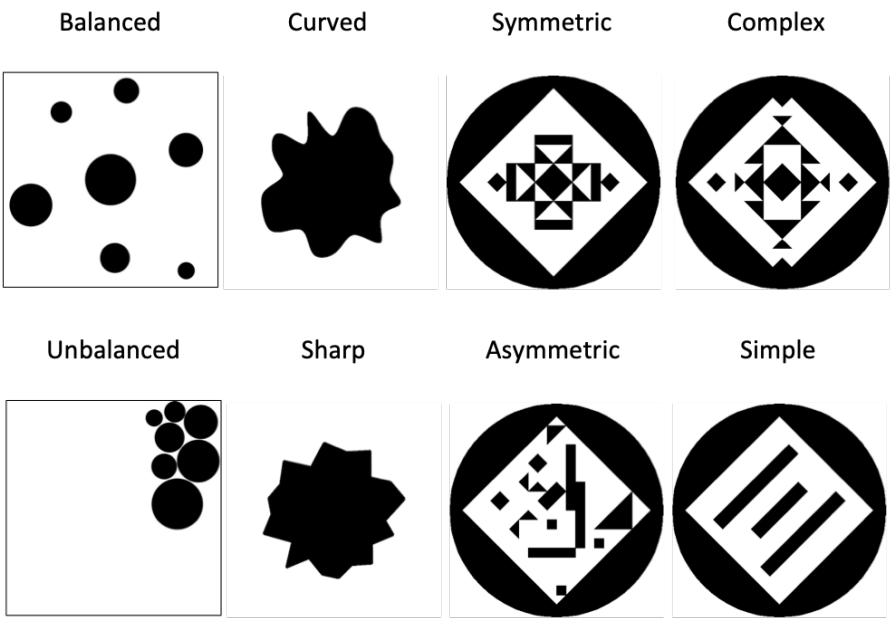


Figure 2



Figure 3

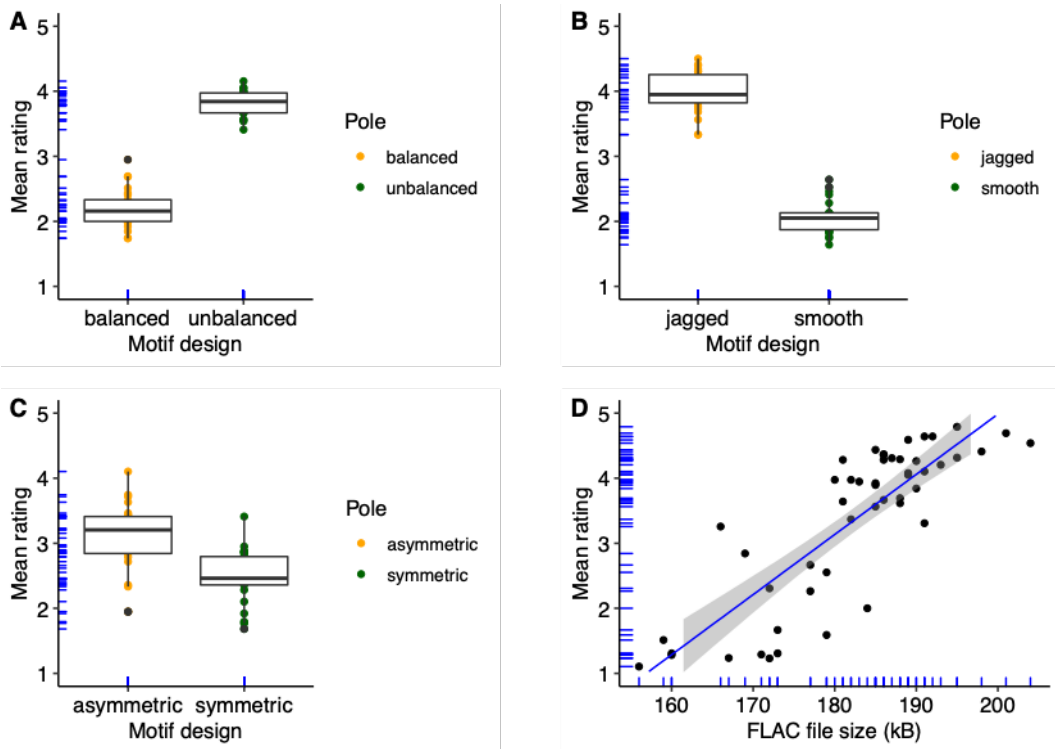


Figure 4

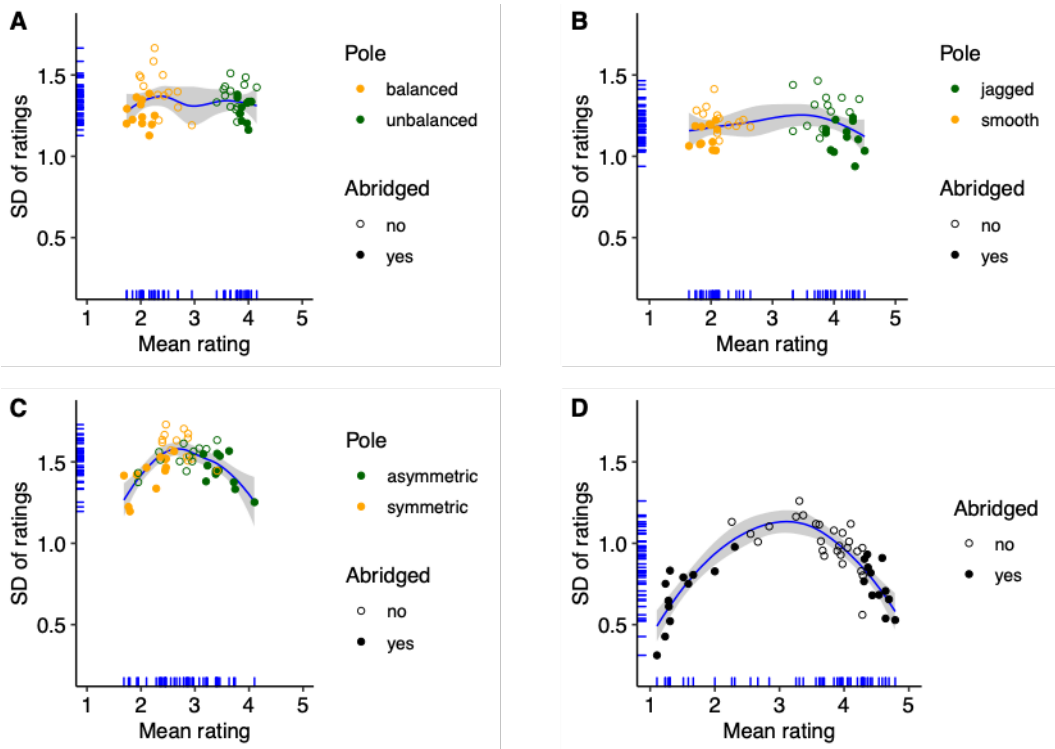


Figure 1. Examples of visual stimuli designed by Wilson & Chatterjee (2005) for balance; Bertamini, Palumbo, Gheorghes, & Galatsidas (2016) for contour; and Jacobsen & Höfel (2002) for symmetry and complexity.

Figure 2. Musical stimuli sample scores in each subset, all to be played in $\text{♩} = 120$ (i.e., quarter note at 120 bpm).

Figure 3. Correspondence between the behavioral assessment and the design of the motifs. Boxplots are used for the discrete subsets of Balance (A), Contour (B), Symmetry (C), and a scatterplot illustrates the continuous subset: Complexity (D). The boxes represent the median, first and third quartiles; whiskers span $Q1 - 1.5 \times \text{IQR}$ (interquartile range) to $Q3 + 1.5 \times \text{IQR}$. For the Complexity subset (D), the regression line is depicted with its 95% CI (gray ribbon). kB refers to kilobytes. The figure includes rug plots of mean ratings (left), and FLAC file size for the Complexity subset (bottom).

Figure 4. Distribution of means and standard deviations of ratings for each musical motif in each subset: Balance (A), Contour (B), Symmetry (C), and Complexity (D). Filled dots correspond to motifs selected for the abridged set. The figure includes rug plots of the mean (bottom) and the standard deviation (*SD*) of the ratings (left). Curved lines depict local polynomial regression fitting ($\text{SD ratings} \sim M \text{ ratings}$), for which the gray ribbon represents the 95% CI.